Cambridge Forum on AI: Culture and Society (2026), **2**, e25, 1–21
doi:10.1017/cfc.2026.10031

REVIEW ARTICLE

Missing context: a review of natural language processing for archives

Lavinia Dunagan, Jesse Johnston and Dallas Card

School of Information, University of Michigan, Ann Arbor, MI, USA
Corresponding author: Lavinia Dunagan; Email: laviniad@umich.edu

(Received 28 February 2025 UTC; revised 11 May 2026 UTC; accepted 23 June 2026 UTC)

Abstract

Archives are essential vehicles for both historical research and memory practices like genealogy. The sheer volume of documents produced via digital infrastructures, however, has posed a challenge for traditional archival processes that rely on human judgments. Some archivists have responded to this challenge by engaging with research in natural language processing (NLP), a subfield of AI that offers practical tools for working with large corpora of text. In this paper, we review applications of NLP methods in archival processes in order to surface fundamental tensions between archival values and automation. NLP can be used to classify materials as sensitive, recognize names in documents and facilitate new kinds of search, among other applications. Our discussion of these issues follows the journey of a document through an archive, from acquisition to interpretation. We find that *sensitivity to context, transparency, resource limitations* and *maintenance* appear as recurring challenges that are complicated by automation. The benefits that NLP technologies lend to managing large archives, however, suggest that these tools should be considered, not discarded. Turning to a more forward-looking discussion of generative AI and archival infrastructure, we argue that the narrative and dialogic potential of contemporary large language models has an especially contentious relationship with conventional archival desiderata.

Keywords: archives; infrastructure; natural language processing

1. Introduction

Over the past century, archives have faced an ever-growing problem of scale, with processing backlogs and record volume far outstripping both institutional resources and available tools (Greene and Meissner 2005; Trace 2022). Archivists have thus turned to tools from natural language processing (NLP) – a subfield of artificial intelligence which deals with human language and text – in the hopes of improving automated processes and access, especially for born-digital or digitized documents (Hutchinson 2020). Most recently, advances in generative AI have created new opportunities and challenges related to whether, where and how this technology should be used in archival settings. These recent developments can be best understood against the broader history of computational methods in archives. In order to understand how greater use of NLP and generative language technologies specifically may impact archives, we review the literature documenting how NLP has been applied to archival collections, surfacing both motivations and challenges that transcend individual systems. Having grounded ourselves in existing applications, we also speak to how large language models (LLMs) may be used in comparable ways and in one entirely novel way: the generation of

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

fluent and appropriate descriptive text with respect to a collection or record. As such, this review is relevant to archivists and scholars of archives who are curious about computational methods, as well as computer scientists wanting to apply their work to archival collections.

In brief, NLP systems can be used variously to classify materials as sensitive or not, recognize text in documents, improve search interfaces for archives and produce metadata. To illustrate these issues, we follow the journey of a document through fundamental archival processes, from acquisition to interpretation. We differ from several previous reviews of AI in archives (Colavizza et al. 2022; Mannheimer et al. 2024; Rolan et al. 2019) in that we are exclusively focused on systems that process, represent and manipulate text, and not computer vision or other non-linguistic applications of machine learning. We also prioritize the thematic continuities between older applications and nascent uses of LLMs in archives. This leads us to observe recurring challenges in even analog processing that may be exacerbated by automation through NLP: *sensitivity to context, transparency, resource limitations* and *maintenance*. At the same time, the benefits NLP technologies lend to managing large archives – providing greater access, interpreting collections in new ways and situating documents in different contexts – render them indispensable.

More recently, NLP has undergone a seismic shift in both capabilities and uptake, thanks to the rise of LLMs. This review emphasizes the continuity between LLMs and older approaches; considering the legacies of previous computational work in archives is informative about both enduring challenges and potential second-order impacts of LLMs. Reliably accounting for the social context of a record in particular presents epistemic questions in addition to technical challenges. While many uses of generative AI are similar to past applications of NLP, LLMs also introduce truly novel possibilities. Turning to more speculative considerations, we argue that the potential use of generative models to create free-text descriptions and dialogue-based archival interfaces has an uneasy relationship with both the historical-evidentiary role of the archive and conventional archival principles.

2. Background

To provide the necessary context for this work, we first briefly describe and summarize the core concerns of both archives and NLP, and then describe in more detail the scope of this review.

2.1. Archives

Our journey begins by addressing archives from the perspective of archival studies. Archives are generally understood as institutional bodies that collect, preserve and provide access to collections of records of historical interest (Prescott and Wiggins 2023). They undergird academic, governmental and community practices of collective memory construction. As Millar (2017) summarizes, "[a]rchives are tools that people use to look beyond the present moment and understand the wider context of a family, community or society." Technical problems in archives can be separated into three processing stages for our purposes. To simplify somewhat, these stages involve (i) acquisition and appraisal; (ii) processing of collections, specifically the description of records and the production of metadata; and (iii) provision of access to a collection through interfaces (Moss and Gollins 2017). Appraisal entails evaluating whether a given document or set of documents should be included in an archive. Description, meanwhile, involves metadata creation and assessing materials for a number of attributes. Archival interfaces entail systems – useful for both archivists and archive users – for interacting with a given collection.

Modern archival practice originated in the state archives that emerged out of the creation and rationalization of various state bureaucracies during the 19th century (Cook 2013). The two most influential principles from this period, which continue to shape archival practice, are *original order* (the maintenance of organizational schemes developed while records were in active use) and *respect des fonds* (the maintenance of records grouped according to their originator). These principles

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

produce an enduring normative assumption in archival work: reorganization of records should be minimized (Nesmith 2002; Yeo 2012). At the same time, archives are also theorized as sites at which collective memory is constructed and renegotiated (Schwartz and Cook 2002). Contemporary archives contend with a fundamental tension between the deeply interpretive work that characterizes historical meaning-making and the seemingly infinite stream of documents produced by modern industrialized life. Understanding the implications of applying NLP to archives requires an account of this conflict, driven by the underlying transparency-, context-, resource- and maintenance-related concerns mentioned above.

Archival work in the 21st century is characterized by additional theoretical and practical desiderata that archivists may seek to balance with these principles. As described by Cook (2013), the move away from an early focus on preserving all evidence of organizational activity was precipitated by the corresponding bureaucracies' growth and the novel requirement in the early 20th century that archives choose which documents to preserve. The fundamental challenge archives are faced with now is the many, many documents that the 21st-century bureaucracy or online platform produces (Thibodeau 2009). Newer frameworks engage directly with this problem of scale. MPLP, or "More Product, Less Process" (Greene and Meissner 2005), is an approach to managing archival backlogs that prioritizes making materials available in great volume, in contrast with other goals like meticulous description. The applications discussed in this paper are motivated by the immense challenges of handling very large collections with limited resources. We note that some of these challenges are due to problems beyond the scope of the review, such as labor shortages due to poor pay (Bredbenner et al. 2024), funding contingent on the political environment (Harper et al. 2024), reasonable distrust in computational methods (Jaillant and Rees 2023), and the difficulty of digitization (Archives 2016; Tudico 2022).[1]

Based on our review criteria, most papers we discuss focus on government archives, likely because they have the resources needed to explore new ways of processing documents. National archives in particular are subject to both the voluminous digital exhaust of bureaucracy and culturally subjective notions of national identity and historical importance (Martínez-Cardama and Pacios 2022). While such institutions are not necessarily representative of typical archives (which run the gamut from small private collections to massive corporate holdings), the techniques we discuss nevertheless have wide relevance – especially as the technology involved becomes more mature.

2.2. Natural language processing

NLP broadly refers to the use of computation to understand – segment, label, reorganize, aggregate and interpret – natural language. The field can be understood as an outgrowth of applied machine learning and computational linguistics (Eisenstein 2019).[2] Modern NLP is itself grounded in large-scale data curation efforts, from the Penn Treebank (Marcus 1993), a dataset of Wall Street Journal articles with part-of-speech labels and other syntactic information numbering in the millions of words, to Common Crawl (Common Crawl – CC-MAIN-2025-08 2025), a collection of billions of scraped web pages that is the foundation for many LLM pretraining projects (Baack 2024). With respect to archives, three practices emerge as particularly relevant: supervised text classification, unsupervised representation learning and information retrieval. Supervised text classification entails training a model to predict labels (or classes) from text, typically based on human-labeled training data, and has natural applications to archival tasks like sensitivity review (Baron et al. 2022), which entail sorting records into predefined categories. Unsupervised approaches, such as topic modeling (Blei et al. 2003), produce interpretable vector representations of documents or collections, often

[1]For example, as of 2025, the U.S. National Archives reports it has digitized about 445 million pages, just 3% of its estimated 13.5 billion physical pages.

[2]Machine learning itself is concerned with creating "computer programs that automatically improve with experience," usually statistical models that infer probabilistic rules from data (Mitchell 1997).

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

based on internal patterns like the co-occurrence of words. Although information retrieval as a field extends beyond NLP, many of its representational techniques originate in language technologies. In particular, search is an umbrella category with functions that range from finding exact terms in documents to using contextual representations of text to retrieve semantically similar documents (Jurafsky and Martin 2023).

While these three families of approaches (as well as other NLP tasks) have historically been pursued somewhat separately, contemporary NLP involves work that is premised on more task-agnostic models – a reality made possible first by more sophisticated word embedding models like Word2Vec (Mikolov et al. 2013) and then by the transformer architecture (Vaswani et al. 2017) and its use in LLMs (Raffel et al. 2020). This type of model was popularized in the late 2010s in the form of BERT (Devlin et al. 2019) and GPT-2 (Radford et al. 2019) and is usually trained to predict the next token given some context using massive text datasets.[3] While the task of predicting the next token is not necessarily of interest in itself (except insofar as it allows for *generating* text), the resulting models have great value as encoders. That is, they transform text into representations that can be used for many potential applications. In addition, many models exhibit the ability to perform a wide range of tasks out of the box without additional training data (so-called zero-shot performance), although rigorous evaluation of many of these capabilities remains difficult given their context-dependent and interpretive nature. Archival processing tasks in particular may be difficult to represent through a traditional NLP evaluation method like benchmarking (Raji et al. 2021).

Contemporary transformer-based NLP systems are technically capable of incorporating much more additional context, but in practice still only consider limited knowledge of the user (Soni et al. 2024). They tend to be data- and energy-hungry (Hedderich et al. 2021; Strubell et al. 2019), opaque (Fanconi et al. 2023) and difficult to repair or update (De Cao et al. 2021). These qualities are at odds with the nature of archival infrastructure – context-driven, sometimes lacking in resources, ideally relatively transparent and composed of many interlocking schemas and standards.[4] This mismatch means that NLP's great promise for managing large archives meets some evident obstacles, although these tensions can be negotiated through both technical strategies within NLP and creative use of NLP by archives.

2.3. Review methodology

To scope this review, we focused on published academic work, located both within the field of archives and in the NLP literature. An initial set of papers from NLP was found by searching the Association for Computational Linguistics Anthology (aclanthology.org) for the word "archive." Additional related papers were drawn from Association for Computing Machinery venues such as the *Journal of Computing and Cultural Heritage*. We also searched *The American Archivist, Archivaria* and *Archival Science* for "NLP" and "natural language processing." Most of the application papers found through this process are case studies. Because the application of generative AI to archives is still nascent, we did not carry out a full systematic review; we instead highlight recent work for the purpose of discussing emerging trends.

3. A document's path

While there has been considerable work on using NLP to analyze large bodies of text in general, the use of NLP for internal archival applications (to assist *archivists* instead of *archive users*) has remained a relatively niche interest. The earliest paper we noted that explicitly broaches the subject

[3]State-of-the-art conversational models involve a number of other development stages, including tuning the pre-trained model on curated data to encourage certain types of responses (Ouyang et al. 2022).

[4]Here, "context" refers to the primary factors external to the content of a record that impact its meaning (Burke 2022). Section elaborates on this issue.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

of NLP and archives is Greenberg (1998). Greenberg emphasizes NLP's potential for reducing the amount of busywork required of archivists while cautioning against its potentially decontextualizing use in interfaces. She also frames automation as epistemically valuable because it requires archivists to systematize their processes. Although the article itself predates much of contemporary NLP, it asks questions with respect to automated judgments of archival materials that remain relevant today. This first account already identifies NLP systems' insensitivity to context as problematic for archival applications; because the article was written at a time when NLP methods were relatively interpretable and did not always require training data or computing power, transparency and resource constraints are not addressed.

The existing applications of NLP to archives vary between custom-built systems and prefabricated tools and packages. Many of the application papers discussed below consist of case studies – unlike most other research in NLP. Fortunately, the modular nature of the abstracted archival pipeline is amenable to both the nature of software and an understanding of these case studies in terms of which archival processes they automate. The following sections are organized in terms of the path one document might take through an NLP-assisted archival processing pipeline. This pipeline maps on to both traditional conceptions of the typical functions of an archive (see Moss and Gollins 2017) and the structure of the more theoretical concerns enumerated above.

3.1. Appraisal

The first stages of a document's life in an archive consist of acquisition and *appraisal* of whether the document should be included in the archive. Some archives take a positive approach of seeking out materials, while others instead define a stream by which they come to own documents (Schmidt 2011). In many contexts, however, archives' central goal of preservation is complicated by the endlessly proliferating streams of text which characterize the modern world. Resource constraints and massive backlogs can additionally limit archives' processing capacities, thus making automated appraisal desirable. Automated processes have particular utility for managing acquisition of born-digital documents, which are usually much more numerous than a typical analog collection (Jaillant 2022).

The acquisition process in a given archive is influenced by its goals (Cook 2011). An archive that collects *all* documents produced by an organization need not make appraisal decisions based on content; modern appraisal originates in the 20th century, when the volume of documents produced by state and corporate bureaucracies massively increased (Buckland 2017a; McKemmish et al. 2005). Heuristics have been used to constrain the amount of documents ingested into state archives. For example, one type of policy for appraising personal communication entails selection based on the seniority of the sender or recipient, such as the U.S. National Archives' CAPSTONE procedure (Jaillant and Rees 2023). This standard necessarily results in an archive stratified according to the power of candidate documents' authors; it also produces series that chiefly correspond to author identity. Rules of this kind can also have legal ramifications, e.g., if messages with evidentiary content are not recognized as such and then not acquisitioned into an archival collection (Johnston et al. 2022; Shilling 2006). NLP offers the potential to more flexibly make similar kinds of automated decisions, but nevertheless involves additional challenges.

In an inventory of prior work on computer-assisted appraisal, C. A. Lee (2018) describes a dominant paradigm of (1) using software to filter digital documents via heuristics in metadata, such as time of creation, and (2) *manually* choosing messages based on content. This is grounded in concerns about transparency and context sensitivity; complete automation of appraisal is not preferred because it is opaque, and it lacks the contextual interpretative capabilities of a human archivist. Prefiguring other papers in this review, Lee's article focuses on digital records generated by large organizations, like governments. Digitization is costly. Thus, digitally automated appraisal is only reasonable for

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

documents that are either born-digital or made readable through digitization and optical character recognition (OCR) prior to an archive attempting to collect them.

NLP can be used to automate complex and subjective judgments about the content of documents, so long as there is available data representing analogous judgments. A major content-based desideratum for many archives is some notion of historical importance (Cox and Samuels 1988). The possibility of using automated text classification for communication archives in particular, given an operationalization of this notion, has attracted attention from multiple domains (Schneider et al. 2019; Vellino and Alberts 2016). For example, Risi et al. (2019) attempt to assess how well future historical relevance can be predicted contemporaneously using about two million cables sent within the U.S. Department of State from 1973 to 1979. A very small subset of these documents was later included in a collection constructed by State Department historians, Foreign Relations of the United States (FRUS), that reflects pivotal moments in U.S. foreign policy. Because it is more carefully prepared for consumption by researchers than the rest of the cables, meaning its content takes on greater authority in the context of computational studies that rely on clean data; as Jo (2020) shows, studying FRUS at scale with computational techniques evinces major shifts in the ideologies of American diplomats. The theoretical problems Risi et al. enumerate range from the relationship between an event's historical relevance and events that have not happened yet to a lack of clarity about the relationship between discrete *events* and *documents*. Using document representations from a topic model as input to a classifier, they achieve an F1 score of 0.86 on a balanced test set of FRUS and non-FRUS cables – an impressive score, but likely not usable in a real archival system.[5] They find that historical importance is not easily predicted by contemporaneous assessments, casting doubt on the feasibility of using that information in an automated fashion.[6]

Venkata et al. (2022) are similarly concerned with government documents, albeit at the British National Archives. In 2018, the volume of digital materials belonging to the British government was estimated at "more than 16 billion emails and 3 billion documents" (Ranade 2021). Venkata et al. conduct a survey of open-source or commercially available AI-based methods for appraising this massive number of documents. They are interested in training classifiers using prior appraisal decisions; these appraisal decisions are themselves grounded in the notion of "enduring value" to the public (although the operationalization of this notion varies). In some settings, documents are classified as retained or not; in others, they are assigned a specific "retention schedule" score corresponding to a degree of default postprocessing. Based on their experiments, the authors observe that no system is satisfactory on its own – i.e., using them would require a human in the loop.

Although NLP models typically classify each document individually, Rolan et al. (2019) question whether the historical relevance of a document can be predicted in isolation. An item-level approach to appraisal misses the contextual importance of documents that gain their meaning from others. Work in government archives also speaks to the additional labor required to use a model within a preexisting standard. As we discuss in Section, descriptive frameworks can be both difficult to operationalize and specific to a given archive. While NLP seems potentially well suited to assisting with the automation of appraisal decisions, the literature indicates mixed results. The core problem that Shabou et al. (2020) identify is in producing a "defensible" method which aligns with archival norms. Having experimented with automation in a Swiss government archive, the authors conclude that automating appraisal requires "prioritizing the implementation of metrics with high automation potential" – an approach which may prejudice archives toward the use of simpler and less seemingly subjective variables in deciding which documents are worthy. Shabou et al. ultimately produce a proof-of-concept tool for manual appraisal which incorporates automated

[5]The F1 score is the harmonic mean of precision (the proportion of predicted positives that are truly positive) and recall (the proportion of true positives that are predicted as positive).

[6]In one project on gray literature in public health (Turner et al. 2002), an analogous situation is observed: some research only becomes relevant to a field after years or even decades.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

scores related to a given document (e.g., topic) without giving decision-making wholly over to the machine.

Notably, many archives' initial processing includes some form of privacy review (Yoxall 1984). This entails the review of records for information that may be inappropriate for public access; it is frequently discussed in the context of the Freedom of Information Act, which mandates the release of government information that is not sensitive when requested (Windon and Youngblood 2024). Technical paradigms discussed above, and supervised classification in particular, have been used to partially or fully automate this process (Baron et al. 2022; Giannella et al. 2024; McDonald 2019; McDonald et al. 2014). Even journalistic archives face related problems with recognizing the copyright status of poorly documented records (Taurino et al. 2025). In general, detection of sensitive information is an exemplary problem in that it is culturally contingent, dependent on context (Nissenbaum 2011) and sometimes in tension with other archival desiderata.

3.2. Description

In addition to classification decisions, NLP also lends itself to the production of metadata. In many archives, the primary interface through which materials are presented to users is the *finding aid*. This is a document that describes the contents of a particular collection, series or even individual record (Chung and Irwin 2017). It is used both as a proxy for the relevant material and as a way of providing context, usually in the form of narrative text (Wiedeman 2019). Usability, satisfaction of formal requirements (such as those of common formats, like the Encoded Archival Description XML standard), and the creation of (contextually) helpful unstructured descriptions are all important desiderata (Berner and Haller 1984; Cox 2008). Similarly, metadata plays a role in the representational schemes discussed below – it can be a mechanism for users to find records – and in description (Buckland 2017b).

In existing work using NLP, we find more discussion of the creation of new types of metadata than of integration with preexisting schema. One straightforward way to extract descriptive information from archival documents is named entity recognition (NER), the labeling of sequences in a text as referring to a certain person, place or thing with a proper name (Daniel and Martin 2023). As described by Ehrmann et al. (2023), a survey of NER in archives, the ability to search a collection for named entities is desirable for many users whose historical work may be focused on a set of individuals or institutions. As Schneider and Chan (2016) put it, NER is a natural tool for addressing two major goals of archivists: "screening for confidential, restricted, or legally protected information … and preparing the files for use by researchers, including by efficiently optimizing discovery and access to the material's intellectual content." Although NER models are intended to generalize, they are likely to perform best on entities included in their training data (Agarwal et al. 2021). As such, performance in a novel archival setting might be worse without additional effort put into labeling additional documents and retraining an NER model.

Production of useful metadata for multimodal and non-digitized archives is another popular application of NLP. Non-digitized archives with preexisting finding aids still contain text on which one can run NER. This is the case in Wright et al. (2023), where a search engine was built for an archaeological archive using NER; their finding aids already documented location, time period, etc. More radically, non-digitized text archives can be made accessible with computer vision models designed to recover a specific part of a document's textual content, such as the U.S. Census race label recognizer described in Jansen and Marciano (2025) for research on Japanese internment or the card classifier used in B. C. G. Lee (2018a) to retrieve death certificates in a large Holocaust archive (both of which do not require OCR). Archives with audio data also tend to have textual finding aids which are possible inputs to NLP-based interfaces themselves; however, full description of the materials may require transcription of that audio.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

In contrast to the content-oriented focus above, Davet et al. (2023) explored the inclusion of *paradata* – here, the byproducts of automated archival pipelines – in metadata. Emphasizing the different questions that different users might have, they recommend including information about the data used to train and validate any models used, as well as their version and performance on representative data. The quirks of a system can have a significant impact on derived metadata; indeed, Liu et al. (2025) creatively use the failings of an NLP tool – Google's language identification application programming interface (API), which struggles with indigenous languages – to find records of indigenous culture in colonial archives. Given that the content of training data may have unpredictable impacts on a model's application to a different domain (Elsahar and Gallé 2019), providing details about that data also constitutes an attempt to provide context. Notably, this concern about a lack of transparency in archival description also applies to completely analog archives; as Douglas (2016) argues, a collection is influenced by archiving practices and "interested parties" (e.g., the subject of a personal collection) before it is even ingested into an archive.

3.3. Interfaces

The question of *how* one should order and present archival collections is one of the central focuses of archival theory – as the paradigmatic principle of original order ensures. The local context of materials impacts their meaning, whether on a web page or in a physical space. For example, a diary may be grouped under materials related to a certain marginalized community, rather than as part of a personal collection. This means the diary will necessarily be interpreted differently – likely as an exemplar of the marginalized group's history, rather than as a document of an individual experience (Caswell et al. 2017; Gabiola et al. 2022). The historiographic responsibility of archivists first detailed a century ago and its more recent evolution into an admission that archivists will always be implicitly supporting one interpretive frame (Cook 2013) has significant implications for computational processes that present archival content to users (Trace 2021). What seems to be the appropriate context for a document is always itself changing in response to shifts in the normative belief systems of the surrounding community (Caswell et al. 2017). In response to the historiographical imperative and increasing recognition of the community impact of archives, many archives have made efforts to make their collections more accessible (Caswell et al. 2016). A core element of this effort has involved implementing search engines and other exploratory tools that support traversal of collections' contents. This subject already has a great deal of overlap with the applications discussed below because the family of technologies supporting many archives' search engines – information retrieval systems – partially constitute a subfield of NLP.

Reindexing of records, contra original order, is sometimes ideal for specific communities' needs. For instance, interfaces that rely on NER, not physical collocation, are one option that can be especially useful for academic users. This is likely because search of NER labels is an effective and relatively efficient way to find documents related to a given historical figure or organization (Schneider and Chan 2016). One example of this is Handler et al. (2022), a platform for exploring a newspaper archive. That project uses NER to index the archive and presents results returned via three separate interfaces for different contexts in which one might view the query (a system built in response to journalist interviews). This exemplifies how archival desiderata considered "typical" may not be a priority, depending on the archive's domain and its expected users. Similarly, both Ranade (2016) and Hinrichs et al. (2015) construct record-level indices for individuals and locations (respectively) because they are sensible axes along which to explore the archives in question – the British National Archives and a corpus of 19th-century commodity trading documents. In general, bespoke search systems that incorporate NLP need not be opaque or context-agnostic; in fact, NLP can be used to extract features of particular interest to a specific user audience.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

The contextual nature of a user's search needs can be addressed in other ways. Topic modeling is a means of categorizing a corpus of documents according to emergent themes; it constitutes an interface that might assist historians and others interested in topical through lines (Muresan and Harper 2004). In existing work on archives, however, topic modeling is frequently a first-line approach for producing record-level description (Coeckelbergs and Van Hooland 2020; Hengchen et al. 2016) and thus only conducted once, in order to produce a content index. These approaches rely on emergent patterns in the corpora they describe, rather than descriptors chosen by archivists. As Underwood (2014) puts it, fitting a topic model entails "harmonizing … initial human assumptions with the complex evidence presented by the documents themselves" in a way that makes those assumptions explicit.

Automating processes related to appraisal and the creation of metadata primarily implicates decisions by archivists. By contrast, interfaces inherently involve decisions being made by users. As Hedstrom (2002) argues, the advent of digital archival interfaces both reminds us that archival infrastructure constrains interpretation and that the typical interfaces an archive provides to a user, such as finding aids, impose a "linear narrative structure" on the contents of a collection – sometimes obscuring discontinuities and gaps. Archivists are obligated to provide context for the records in a given collection; this context ideally includes the construction of the archive itself. Moreover, the histories of digital records are important parts of their information content. Acker (2017) suggests that digital records are characterized more by the traces of their mutation across different platforms than even their content; archives' power over the evidentiary status of records depends on recognition of this quality. Automating archival processes, then, requires attention to both the modifications made on the metadata paratext associated with a record and the "exhaust" created by the processes themselves, like the *paradata* that Davet et al. (2023) refer to.

4. Generative AI along the path

The case studies above all make use of relatively standard methods in NLP – that is, ones for which relative consensus has been reached on appropriate datasets and evaluation methods, and for which some precedent exists in at least a decade of computational research. The shift in focus to LLMs since the release of ChatGPT in 2022, however, has resulted in an eruption of literature about how text generation models might be used to replace both humans and older NLP methods in various socially important processes. At the same time, few standards exist for evaluating the performance of these systems in specific archives. Validating that a model's output is appropriate for a given context, likely users, and the remit of the archive itself is not easily done via evaluation on benchmarks, as noted by Toth et al. (2025). The papers discussed below do not conduct systematic evaluation of their applications with both users and relevant performance metrics.

The major capability of LLMs that differentiates their potential use in archives from previous applications of NLP is the *generation of fluent and contextually relevant text*. The flexible nature of LLMs means that they can also easily be adapted to the processes discussed above, such as text classification, but it is in the ability to generate fluent text that the most transformative potential of LLMs has been observed. As of yet, there is comparatively little published empirical work on the effects of LLMs on typical archival pipelines. Nonetheless, we can speculate about thematic concerns, especially in the context of prior work on NLP systems and interpretation (Nelson 2022; Shah and Bender 2022). How might LLMs be used in a manner that respects the evidentiary status of archival documents themselves? What new ways of interfacing with archives might LLMs make possible? To rephrase Greenberg (1998), how can we reduce the impact of LLMs' weaknesses while exploiting their

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

strengths? While none of these questions are fully answered below, we hope to frame LLMs' generative capabilities in archives in terms of related work in adjacent fields and theoretical commentary on archives and interpretation.

4.1. Generative AI in archival infrastructure

Numerous organizations have begun to explore the use of generative AI, such as LLMs, to describe and index records. The exact features entailed vary with the archive. For instance, Potter and Saccucci (2024) – a project exploring LLM-based indexing at the Library of Congress (LOC) – is concerned with producing standardized Machine-Readable Cataloging (MARC) descriptions in addition to specific subject headings.[7] In this section, we review the LOC project as well as three *generative* uses of generative AI: a framework for using LLMs to interface with web archives, an article describing the use of LLMs to produce more comprehensive metadata for the Grand Ducal Archive of the Medici, and an article presenting an LLM and its uses in the Chinese state archive from which its pretraining data came. The second use exemplifies how LLMs may be used in contexts where traditional approaches to processing and interface construction may be inappropriate, while the third and fourth demonstrate the use of an LLM to automate a descriptive process in a more classical archival setting. We avoid reviewing papers that do not explicitly discuss a particular archive, such as Nguyen et al. (2025), or that deal with personal collections, such as Garcia and Weilbach (2023), although these works also suggest possible uses of LLMs in archives.

4.1.1. Ebook metadata

In an ongoing project titled "Towards Piloting Computational Description" (Potter and Saccucci 2024), the Library of Congress describes experimenting with using a fine-tuned LLM to produce metadata for ebooks (currently with the expectation of human review). This project exemplifies the type of additional work required to adapt a tool like LLMs, which are presented at times as universal. The authors fine-tuned Mistral AI's mixture-of-experts model, which has 46.7 billion parameters in total (Jiang et al. 2024), to produce different versions of MARC with varying levels of explication of metadata fields. Given the ebook's full text, the model excelled at extracting values for formally discrete elements like the title, but struggled to produce appropriate subject headings from the LOC's own categorization system (Library of Congress 2025). This system consists of tens of thousands of hierarchically organized subjects (e.g., "United States – History – Periodicals" or "Lullabies, Urdu"); the model variously mislabeled documents with headings that were too narrow, too broad or nonexistent. The Library turned to an embedding-based method that searched for the subject headings of semantically similar books that had already been cataloged. LLMs excel at generating fluent text and copying relevant text from unstructured input, but may not yield the most parsimonious judgments based on a category system with a closed set of labels. The system currently being iterated upon incorporates multiple NLP components with distinct purposes, rather than a "one size fits all" approach.

4.1.2. Web archive chatbot

The scale of the LOC is massive; it is rivaled in both size and heterogeneity, however, by web archives. Web archives tend to aspire to comprehensive storage of an online environment (e.g., all URLs ending in ".gov"), rather than careful selection for relevant content (Ogden et al. 2017; Ogden and Maemura 2021). Construction, description and exploration of web collections have all been framed as novel challenges in the field of archives (Summers and Punzalan 2017). The networked and dynamic nature of digital material writ large and the internet especially challenge traditional assumptions undergirding typical archival processing (Acker 2017). Enabling user exploration of billions of webpages is

[7]MARC is a commonly used record format in archives like the LOC (Library of Congress 2024).

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

difficult – for Google, certainly, but especially for small institutions interested in supporting rigorous historical inquiry. WARC-GPT is a framework for enabling conversational use of a web archive via a retrieval-augmented generation (RAG) (Lewis et al. 2020) system incorporating an LLM of one's choice (Cargnelutti et al. 2024). It uses the LLM to construct a chat interface between the user and the archive. Retrieved documents are synthesized within the interface, although they can be accessed after clicking another pane. The system was evaluated using factual queries; it successfully retrieved documents and accurately characterized events that occurred after the LLM was trained. These last two features speak to ongoing ambiguity surrounding the informational purpose of both LLMs and web archives. Contrasting WARC-GPT with the Internet Archive's Wayback Machine (Internet Archive: Wayback Machine 2001), the authors note that WARC-GPT is not intended to facilitate "exploring and comparing multiple versions of a given page archived repeatedly over time or for performing large-scale analysis of web archives metadata."[8]

4.1.3. Structuring user-generated keywords

LLMs can be especially useful in contexts where users may benefit from the expansion and rearticulation of preexisting free-text descriptions of records. The Grand Ducal Archive of the Medici is a partially digitized resource for historians of early modern Italy, consisting of about 3 million letters and 7 million other documents; all are labeled with one of 44 topical keywords and most have a synopsis provided by employees or users of the archive (The Medici Archive Project 2025). Toth et al. (2025) describe the use of ChatGPT to (1) enrich these topical keywords, (2) create a new interface to the letters via semantic embedding of the synopses and (3) extract relationships and events in structured form. Their approach to keyword expansion is distinguished by its reliance on an interpretability method from NLP (Nielsen et al. 2022) to derive relationships between preexisting topical keywords and synopsis vocabulary – allowing them to remain grounded in the archive's schema. The authors also experiment with using ChatGPT to summarize synopses further and then embed the results. Lastly, ChatGPT is similarly used to extract relationships, events and locations from synopses, producing metadata useful for both summary visualizations and navigation. As the authors acknowledge, all of these tasks are possible with traditional NLP systems. However, limited resources and the abundance of unstructured user-generated descriptive text made an LLM-based approach for additional indexing extremely appealing. In a similar paper, Reusens et al. (2025) claim that the primary problems with using LLMs for metadata generation lie in hallucination and ahistorical framing; furthermore, using a more parsimonious prompt that attempts to avoid these issues can actually decrease performance on the main task. As Toth et al. note, the challenge is in validating their outputs according to the perspective of the archive and its contents – a challenge for which no large-scale archival benchmark exists.

4.1.4. Purpose-built archival LLM

Evaluation of archival systems that incorporate LLMs has been discussed in other contexts – namely, that of state archives. Zhang et al. (2024) train a 7 billion parameter transformer model, ArcGPT, for various archival processing tasks on 100 billion tokens of real administrative records from a Chinese government office. Motivated by a life-cycle perspective on archival records, Zhang et al. focus on four central tasks which relate primarily to appraisal: "retention period prediction, open-access identification, confidentiality prediction, and post-OCR processing." The first three are classification tasks, while the last involves text generation – namely, transforming flawed OCR transcriptions into fluent text. ArcGPT excels at the first two tasks in comparison with other models and attains similar performance to baselines on confidentiality prediction (F1 of 0.94). Nonetheless, postprocessing of OCR output is the application for which ArcGPT is most useful. They also introduce a framework for evaluating their system with users – one which would use interviews to determine if the results from

[8]More recently, others have explored using LLMs for large-scale webpage description, as in the case of Web Archive Singapore (Nair et al. 2025).

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

ArcGPT are compatible with the preferences of archivists, who the authors desire to "emancipate … from the rigors of extensive textual operation."

4.2. Remembering with LLMs

A growing number of exploratory papers have also examined potential uses of LLMs and AI more broadly in memory projects, especially educational ones (see Kouzelis and Spantidi 2024). This work tends to focus on the use of AI to memorialize events which create archival silences (Ghaddar 2016; Hawkes 2012), such as genocides (Makhortykh et al. 2023). Toth (2021), for instance, uses a semantic search system to support the recovery of Holocaust testimonies. In general, however, the use of LLMs in this context remains nascent. Thus, we speculate on the basis of available documentation of reports on ongoing efforts and work on these technologies within NLP.

As discussed above, LLMs can be used to summarize and synthesize text. Moreover, frameworks such as RAG allow one to cheaply integrate an external collection with an existing LLM (Lewis et al. 2020). However, LLMs are not completely faithful or unbiased authors (Bender et al. 2021). GPT-4 has been found to generate descriptions on the basis of demographic information that are stereotypical if not overtly offensive (Cheng et al. 2023). Bias has also been observed in RAG systems, including in historical narratives (Fan and Gardent 2022). Any proposal to use LLMs to generate new text raises questions about appropriateness, fidelity and authenticity.

4.3. Narrative generation

While many potential uses of LLMs closely resemble previous approaches with earlier techniques, the generation of coherent text opens up much more novel possibilities. As such, archival theory has not yet grappled in detail with how to think about and work with such systems. Nevertheless, inspired by recent excitement around "bringing history to life" (Eisikovits 2021), we can look to past debates about the more speculative writing of history, as expressed in the work of Saidiya Hartman and Stephen Best.

Saidiya Hartman's exploration of the limits and possibilities of the archive (Hartman 2008; 2019) provides one means of conceptualizing generative AI in the archives. Hartman's central contribution is "critical fabulation," an approach to interpretivist archival research that entails writing speculative prose about the subjective experiences of marginalized people who are present in the archive but silent. Meanwhile, Stephen Best critiques the expectation that cohesive histories can be recovered from archives (Best 2018), fearing the treatment of the archive simply as "a rich repository of material from which one may recover and reanimate the lives of the enslaved." Best argues that the recuperative stance taken by Hartman entails interpretive irresponsibility: grafting contemporary expectations and norms onto the experiences of people with very different self-conceptions. While we cannot do justice to their work here, the debate between Hartman and Best provides a useful touchstone for thinking about the emerging possibilities and dangers around the use of LLMs in archives.

In principle, generative AI could be used to try to generate the kinds of first-person testimonies that are absent from the historical record for many marginalized groups. It also constitutes a new way of producing cohesive histories out of archives; LLMs can be used to smoothly aggregate textual information and induce the appearance of continuity. In light of Hartman's work and Best's critique, it may be worth considering (1) what kinds of recovery and reinterpretation are made possible by the use of LLMs and (2) whether these processes are still accountable to the histories of discontinuity and disruption that Best identifies. Any use of LLM-generated text in archives, *especially* in the context of perspectives that have been silent, risks the imposition of contemporary perspectives and desires (including those encoded in the LLM) onto historical materials. It will also have to contend with the nature of LLMs' representations of identity groups, which can often tend toward stereotypes (Wang et al. 2024), as well as the discontinuity of identities salient at one time but not another. This is not

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

to say that LLMs are never useful in an archival context where silences abound. Rather, the use of LLMs to critically fabulate the perspectives of the under- and unrepresented itself requires critical scrutiny.

5. Explicating tensions

Surveying the use of NLP in archives helps to surface tensions that arise from attempting to scale up interpretive tasks in an environment that shapes historical consensus. These tensions are amplified by LLMs. We identify *sensitivity to context, transparency, resource limitations* and *maintenance* as particularly salient obstacles for adoption of NLP in an archival setting. These issues are always borne out in the specifics of each case study, but we explicate below how they arise from foundational characteristics of the dominant paradigm in NLP.

5.1. Sensitivity to context

Context can mean many different things (Burke 2022). We recognize three main categories of archival context: the context that records in a collection give to other records in a collection, the context that archivists endow records with through description, and the context that external information known by users gives to records in a collection. We focus on the first and second aspects of context, given their more obvious relationship to archival processing.

First, there is a basic conflict between the NLP paradigm of individual document classification and the contextually aware mode in which archivists appraise, describe and represent archival materials (Cook 2011; Fintland 2016; Yakel 2003). As Rolan et al. (2019) explicitly note, part of the difficulty of automating archival processes is in approximating the holistic means by which archivists make decisions about individual documents while thinking of collections as aggregate wholes. This situation is exacerbated by the networked and nonlinear nature of contemporary communication technologies (Acker 2017). Construction of a set of reasonable "series" out of, e.g., millions of social media posts is far from straightforward.[9]

Second, the choice to endow an archive with one context or another can significantly impact user interpretations as well. Douglas (2016) emphasizes that the traditional mode in which finding aids' descriptive text is written is that of "the view from nowhere" – obscuring the process by which a collection was constructed while narrativizing a seemingly static and universal vision of a given collection's history. This is despite the sometimes contentious processes by which materials are acquired. Recalling the earlier discussion of *paradata* (Davet et al. 2023), Douglas might encourage the production of a finding aid wherein the automation of descriptive choices is motivated and explained within the archive's social context.

5.2. Transparency

Most NLP models used today take the form of neural networks involving many layers of nonlinear functions. Language carries both compositional meaning and other types of semantic content (referential, metaphorical, etc.). As a result, it is frequently difficult to attribute outputs to specific elements of a text and almost always difficult to attribute them to a model mechanism. This poses a particular challenge for using NLP in archives, given the emphasis on explicating processing choices to users (Caswell 2021; Douglas 2016). In the context of appraisal specifically, the invisibility of the discarded materials in a typical archival setting further exacerbates this challenge; only the selected records are visible to the user.

[9]For a practical example of the challenges a comprehensive platform archive project can face, see the Library of Congress' Twitter archive. The LOC began its collection of tweets in 2010 by preserving all text posted on the website, but decided seven years later that this was too burdensome; the volume of tweets had grown significantly, many tweets were primarily image-based, and tweets themselves became longer (Osterberg 2017).

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

Bunn (2020) argues that machine learning in archives is best undertaken with a plan for articulating how a given model functions. This means both making basic facts about the model available (e.g., performance on a representative dataset) in a form that a user can understand *and* conducting an analysis of the model using methods from the machine learning subfield of interpretability. Transformers, however, are sufficiently complex that it is difficult both to visualize mechanisms within the model and to simplify them for the benefit of users (Abnar and Zuidema 2020; Antverg and Belinkov 2022). We might also hope to provide some accessible description not just of the process by which an LLM was used, but also of the data used to train that model. (Desai et al. 2024) and Schoenebeck and Conway (2020) have argued that archival thinking and practices are themselves applicable to the data used in machine learning systems. Proper description of archival processes relies on documentation – ideally archival – of the data used to construct those processes.

5.3. Resource limitations

Successful adoption of NLP systems in general and LLMs in particular at a given archival institution relies on a number of material factors. These include the experiences of archivists there, financial constraints with respect to costs like API fees and the nature of the processing backlog. The limitations of available data in historical contexts – especially *labeled* historical data – can similarly impede system development and evaluation, given the data requirements of large neural models (Hedderich et al. 2021). Annotation in a given archive may need to be undertaken by people with a specific experience, such as expertise or community membership, in order to produce representative training data. Indeed, the annotation projects that support frontier LLMs have been critiqued on the basis of exploitative employment models, the injury done to annotators when they contend with harmful content and other ethical concerns (Crawford et al. 2014; Muldoon et al. 2024); these issues are especially salient in the context of archives that may relate to marginalized populations. LLMs appeal to some of the practitioners whose work we review because of a perceived lesser need for expert annotations, although validating their use also requires human judgments. Practical use of LLMs requires either expensive hardware (with open models) or per-use cloud computing or API fees – whereas earlier methods can be run locally on standard computers. Zhang et al. (2024) is somewhat unique because of how it involves developing an archive-specific LLM; outside of state archives with significant resources, pretraining an entirely new model is likely infeasible. However, we recognize the largest factor in archival resource limitations to be time.

Time is a central concern for archives. Many archivists simply do not have the time to acquire skills that are generally required to use NLP tools seamlessly, such as Python programming (Jaillant and Aske 2024; Poole and Todd-Diaz 2022; Rasmussen 2019). There can be a high upfront cost to adoption of NLP systems (Jaillant and Aske 2024) even if they have the potential to speed up processing in the long run. Archivists at state archives especially tend to be occupied by the massive backlogs of unprocessed (and undigitized) materials that have accumulated over decades, absent a plan for addressing them (Greene and Meissner 2005). As Mordell (2019) observes, the practical outcome of many efforts to digitize and "datafy" archives is frequently much time wasted by archivists and few benefits to users. AI is hardly a panacea for some of the most pressing content-related problems facing archives; the limitations of many collections – especially from the perspectives of marginalized communities – originate in preservation and appraisal decisions made a hundred years ago (Jaillant et al. 2025).

5.4. Maintenance

Maintenance entails both the labor that updating infrastructure and schema involves and the high-level ideation that arises from revising archival processing practices. The manifesto (Acker et al.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

2019) produced by the archival professional collective The Maintainers argues that "moments of upgrade and reimagining can also resource and provide justification for maintenance processes." Ideally, adoption of more sophisticated NLP technologies and the new schema they make possible is also an opportunity to participate in maintenance projects. Conversely, Jackson (2014) contends that new technology's adoption relies on a continuous project of repair which articulates that technology in the context of older infrastructure and addresses its breakdown. LLMs may at first seem inherently flexible in a way that makes maintenance less necessary; intuitively, the prompt templates undergirding systems using LLMs may make them easier to understand and modify. However, the polysemic nature of fluent text may obscure the limitations of an LLM-based system (Shah and Bender 2022), both intrinsically and as interfaces with other components of archival processing infrastructure. As seen in Potter and Saccucci (2024), LLMs are also rarely the optimal choice for tasks involving well-defined yet complex schemas. Adapting any NLP system, but LLMs especially, to function within a preexisting ontology requires its own maintenance and articulation work.

6. Conclusion

Archives are sites of historical interpretation and memory creation. They are also tasked with ingesting and preserving very large amounts of text documents and other materials. This engenders multiple conflicts – between transparency and flexibility, between specificity and speed and between hard material constraints and theoretically prescribed goals. The use of NLP can exacerbate these problems, given the field's current focus on LLMs (which are opaque, resource-hungry and difficult to modify), but it also provides a way to manage the scale of modern collections. A given document that enters an archive may be appraised, described and accessed through a search interface in a completely or partially automated manner. In the archives represented here – especially those with state funding and large staffs – archivists have risen to the challenge by adopting NLP methods for their purposes. The technical tools used to accomplish this in the past vary. From bag-of-words logistic regression for appraisal of emails (Schmidt 2011) to transcription and NER for the production of metadata in an audio archive (Verhagen et al. 2022), plenty of NLP approaches have something to offer to archivists.

The uses of NLP we examine raise important questions about the role of automation in archives. As archives unevenly adopt AI into workflows, the challenges they face – maintaining trustworthiness, keeping human judgments in the loop and evaluating models according to subjective and contextually dependent desiderata – recall older problems faced by practitioners, dating back to Jane Greenberg's 1996 paper on NLP and archives. This is not to say older NLP systems and LLMs have identical impacts on archival processing. LLMs are both less transparent and (seemingly) more general-purpose; their ability to generate fluent text that appears human to the reader is sui generis compared to older NLP systems. Recent advances within NLP and the use of best practices indicate the potential for multiple archival goals to be achieved while maintaining a trustworthy system, especially when systems are custom-built for a specific archive and its users. It is exactly because the purposes and uses of archives are so diverse that NLP has much to offer to them. However, the application of NLP to archives requires transparency, attention to resource constraints and above all *attention to context* in both the design of the technology and its function. Generative AI tools – and LLMs especially – can be appealing technologies, but their capacity to produce fluent text does not necessarily correspond to an ability to enact archival values.

Acknowledgements. We are grateful to Justine Zhang for advice and comments on an earlier draft. We would also like to thank members of the Libraries, Archives and Curation research group at the University of Michigan School of Information for their support, including suggestions on a previous version of this work.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

Funding statement. The first author was supported in part by an NSF Graduate Research Fellowship (DGE 2241144). We also received funding from an endowment by the Carnegie Foundation to support publication of this article.

Competing interests. We have no competing interests to declare.

References

Abnar, S., and W. Zuidema. 2020. "Quantifying Attention Flow in Transformers." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Association for Computational Linguistics. Online, July, 4190–7. https://doi.org/10.18653/v1/2020.acl-main.385.

Acker, A. 2017. "When Is a Record?" In *Research in the Archival Multiverse*, edited by Gilliland, Anne J., McKemmish, Sue and Lau, Andrew J., Clayton, Victoria, Australia, 288–323. Monash University Publishing.

Acker, A., H. Arnold, J. Castro, S. Galvan, P. Hswe, J. Meyerson, B. Nowviskie, M. Lassere, D. Olson, M. A. Parsons, A. Russell, L. Vinsel, and D. J. Wright. 2019. "Information Maintenance as a Practice of Care." *Zenodo*. https://doi.org/10.5281/zenodo.3251131.

Agarwal, O., Y. Yang, B. C. Wallace, and A. Nenkova. 2021. "Interpretability Analysis for Named Entity Recognition to Understand System Predictions and How They Can Improve." *Computational Linguistics* 47 (1): 117–40. https://doi.org/10.1162/coli_a_00397.

Antverg, O., and Y. Belinkov. 2022. "On the Pitfalls of Analyzing Individual Neurons in Language Models." *International Conference on Learning Representations*. https://openreview.net/forum?id=8uz0EWPQIMu.

Baack, S. 2024. "A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAccT'24, 2199–208. https://doi.org/10.1145/3630106.3659033.

Baron, J. R., M. F. Sayed, and D. W. Oard. 2022. "Providing More Efficient Access to Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege." *Journal on Computing and Cultural Heritage* 15 (1): 1–19. https://doi.org/10.1145/3481045.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT' 21,(New York, NY, USA), 610–23. https://doi.org/10.1145/3442188.3445922.

Berner, R., and U. Haller. 1984. "Principles of Archival Inventory Construction." *The American Archivist* 47 (2): 134–55. https://doi.org/10.17723/aarc.47.2.n14378125t7t8547.

Best, S. 2018. *None Like Us: Blackness, Belonging, Aesthetic Life*. Durham, North Carolina, USA. Duke University Press. http://www.jstor.org/stable/j.ctv11smf1b.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Buckland, M. 2017a. *Document Theory: An Introduction*. Morepress Books. https://morepress.unizd.hr/books/index.php/press/catalog/view/5/5/53.

Buckland, M. 2017b. "Metadata." In *Information and Society*, edited by Michael Buckland, 111–134. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/10922.003.0010.

Bunn, J. 2020. "Working in Contexts for Which Transparency is Important." *Records Management Journal* 30 (2): 143–53. https://doi.org/10.1108/RMJ-08-2019-0038.

Burke, P. 2022. "Context in Context." *Common Knowledge* 28 (1): 11–40. https://doi.org/10.1215/0961754X-9713549.

Cargnelutti, M., K. Mukk, and C. Stanton. 2024. "WARC-GPT: An Open-Source Tool for Exploring Web Archives Using AI | Library Innovation Lab." *Library Innovation Lab*. https://lil.law.harvard.edu/blog/2024/02/12/warc-gpt-an-open-source-tool-for-exploring-web-archives-with-ai/.

Caswell, M. 2021. "Dusting for Fingerprints: Introducing Feminist Standpoint Appraisal." *Journal of Critical Library and Information Studies* 3 (22). https://doi.org/10.24242/jclis.v3i2.113.

Caswell, M., M. Cifor, and M. H. Ramirez. 2016. "'To Suddenly Discover Yourself Existing': Uncovering the Impact of Community Archives." *The American Archivist* 79 (1): 56–81. https://doi.org/10.17723/0360-9081.79.1.56.

Caswell, M., A. A. Migoni, N. Geraci, and M. Cifor. 2017. "'To Be Able to Imagine Otherwise': Community Archives and the Importance of Representation." *Archives and Records* 38 (1): 5–26. https://doi.org/10.1080/23257962.2016.1260445.

Cheng, M., E. Durmus, and D. Jurafsky. 2023. "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.84.

Chung, S. K., and K. Irwin. 2017. "Archival Finding Aids." In *Encyclopedia of Library and Information Sciences*, 4th ed., Boca Raton, Florida, USA, 133–140. CRC Press.

Coeckelbergs, M., and S. Van Hooland. 2020. "Concepts in Topics. Using Word Embeddings to Leverage the Outcomes of Topic Modeling for the Exploration of Digitized Archival Collections." In *Data and Information in Online Environments*, edited by R. Mugnaini, New York City, New York, USA, 41–52. Springer International Publishing. https://doi.org/10.1007/978-3-030-50072-6_4.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

Colavizza, G., T. Blanke, C. Jeurgens, and J. Noordegraaf. 2022. "Archives and AI: An Overview of Current Debates and Future Perspectives." *Journal on Computing and Cultural Heritage* 15 (1): 1–15. https://doi.org/10.1145/3479010.

2025. ***Common Crawl - CC-MAIN-2025-08***. https://commoncrawl.org/.

Cook, T. 2011. "'We are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future." *Journal of the Society of Archivists* 32 (2): 173–89. https://doi.org/10.1080/00379816.2011.619688.

Cook, T. 2013. "Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms." *Archival Science* 13 (2): 95–120. https://doi.org/10.1007/s10502-012-9180-7.

Cox, R., and H. Samuels. 1988. "The Archivist's First Responsibility: A Research Agenda to Improve the Identification and Retention of Records of Enduring Value." *The American Archivist* 51 (1-2): 28–42. https://doi.org/10.17723/aarc.51.1-2.gkw67424l3344ug8.

Cox, R. J. 2008. "Revisiting the Archival Finding Aid." *Journal of Archival Organization* 5 (4): 5–32. https://doi.org/10.1080/15332740802153245.

Crawford, K., K. Miltner, and M. L. Gray. 2014. "Critiquing Big Data: Politics, Ethics, Epistemology." *International Journal of Communications - Special Section Introduction* 8 (January): 1663–72.

Daniel, J., and J. H. Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed., Hoboken, New Jersey, USA. Prentice Hall PTR.

Davet, J., B. Hamidzadeh, and P. Franks. 2023. "Archivist in the Machine: Paradata for AI-based Automation in the Archives." *Archival Science* 23 (2): 275–95. https://doi.org/10.1007/s10502-023-09408-8.

De Cao, N., W. Aziz, and I. Titov. 2021. "Editing Factual Knowledge in Language Models." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, edited by M.-F. Moens, X. Huang, L. Specia, and S. W.-T. Yih, Punta Cana, Dominican Republic, 6491–6506. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.522.

Desai, M. A., I. V. Pasquetto, A. Z. Jacobs, and D. Card. 2024. "An Archival Perspective on Pretraining Data." *Patterns* 5 (4): 100966. https://doi.org/10.1016/j.patter.2024.100966.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by J. Burstein, C. Doran, and T. Solorio, Minneapolis, Minnesota, USA, 4171–86. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Douglas, J. 2016. "Toward More Honest Description." *The American Archivist* 79 (1): 26–55. https://doi.org/10.17723/0360-9081.79.1.26.

Ehrmann, M., A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet. 2023. "Named Entity Recognition and Classification in Historical Documents: A Survey." *ACM Computing Surveys* 56 (2): 1–47. https://doi.org/10.1145/3604931.

Eisenstein, J. 2019. *Introduction to Natural Language Processing*, Cambridge, Massachusetts, USA. The MIT Press.

Eisikovits, N. 2021. "The Slippery Slope of Using AI and Deepfakes to Bring History to Life." *The Conversation*. https://theconversation.com/the-slippery-slope-of-using-ai-and-deepfakes-to-bring-history-to-life-166464.

Elsahar, H., and M. Gallé. 2019. "To Annotate or Not? Predicting Performance Drop under Domain Shift." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, edited by K. Inui, J. Jiang, V. Ng, and X. Wan. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1222.

Eun Seo, J. 2020. "Foreign Relations of the United States Series, 1860-1980: A Study in New Archival History." PhD diss, Stanford University. https://www.proquest.com/docview/2457237433/abstract/71D24381FCFC4718PQ/1.

Fan, A., and C. Gardent. 2022. "Generating Biographies on Wikipedia: The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by S. Muresan, P. Nakov, and A. Villavicencio. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.586.

Fanconi, C., M. Vandenhirtz, S. Husmann, and J. Vogt. 2023. "This Reads Like That: Deep Learning for Interpretable Natural Language Processing." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, edited by H. Bouamor, J. Pino, and K. Bali, Singapore, 14067–76. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.869.

Fintland, I. 2016. "Archival Descriptions through the Looking Glass: Paratexts in Wonderland." *The American Archivist* 79 (1): 137–60. https://doi.org/10.17723/0360-9081.79.1.137.

Gabiola, J., G. Brilmyer, M. Caswell, and J. Zavala. 2022. "'It's a Trap': Complicating Representation in Community-Based Archives." *The American Archivist* 85 (1): 60–87. https://doi.org/10.17723/2327-9702-85.1.60.

Garcia, G. G., and C. Weilbach. 2023. "If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History." In *CHR 2023: Computational Humanities Research Conference*, December, 616–38.

Ghaddar, J. J. 2016. "The Spectre in the Archive: Truth, Reconciliation, and Indigenous Archival Memory." *Archivaria* 82 (1): 3–26.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

Giannella, C. R., L. K. Branting, J. A. Van Guilder, and J. R. Baron. 2024. "Improving Automated Detection of FOIA Deliberative Process Privilege Content." *Digital Government: Research and Practice* 5 (2). https://doi.org/10.1145/3651226.

Giulia, T., S. Sweeney, D. Facklam, and D. A. Smith. 2025. "Copyright, Privacy, and Public Access in News Archives: A Proof of Concept on the Boston Globe Photograph Morgue." *AI & SOCIETY* 40 (6): 4363–81. https://doi.org/10.1007/s00146-025-02208-x.

Greenberg, J. 1998. "The Applicability of Natural Language Processing (NLP) to Archival Properties and Objectives." *The American Archivist* 61 (2): 400–25. https://doi.org/10.17723/aarc.61.2.j3p8200745pj34v6.

Greene, M., and D. Meissner. 2005. "More Product, Less Process: Revamping Traditional Archival Processing." *The American Archivist* 68 (2): 208–63. https://doi.org/10.17723/aarc.68.2.c741823776k65863.

Handler, A., N. Mahyar, and B. O'connor. 2022. "ClioQuery: Interactive Query-Oriented Text Analytics for Comprehensive Investigation of Historical News Archives." *ACM Transactions on Interactive Intelligent Systems* 12 (3): 1–49. https://doi.org/10.1145/3524025.

Harper, L., W. Burr, and T. Blanton. 2024. "U.S. National Archives' 2025 Budget Request Threatens Mission Failure." *National Security Archive*. https://nsarchive.gwu.edu/foia-audit/foia/2024-03-15/us-national-archives-2025-budget-request-threatens-mission-failure.

Hartman, S. 2008. "Venus in Two Acts." *Small Axe* 12 (2): 1–14. 10.1215/-12-2-1.

Hartman, S. 2019. *Wayward Lives, Beautiful Experiments: Intimate Histories of Riotous Black Girls, Troublesome Women, and Queer Radicals*. New York, W. W. Norton & Company.

Hawkes, M. 2012. "Containing Testimony: Archiving Loss after Genocide." *Continuum* 26 (6): 935–45. https://doi.org/10.1080/10304312.2012.734600.

Hedderich, M. A., L. Lange, H. Adel, J. Strötgen, and D. Klakow. 2021. "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, 2545–68. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.201.

Hedstrom, M. 2002. "Archives, Memory, and Interfaces with the Past." *Archival Science* 2 (1): 21–43. https://doi.org/10.1023/A:1020800828257.

Hengchen, S., M. Coeckelbergs, S. van Hooland, R. Verborgh, and T. Steiner. 2016. "Exploring Archives with Probabilistic Models: Topic Modelling for the Valorisation of Digitised Archives of the European Commission." In *2016 IEEE International Conference on Big Data (Big Data)*, December, 3245–49. https://doi.org/10.1109/BigData.2016.7840981.

Hinrichs, U., B. Alex, J. Clifford, A. Watson, A. Quigley, E. Klein, and C. M. Coates. 2015. "Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration." *Digital Scholarship in the Humanities* 30 (suppl_1): i50–75. https://doi.org/10.1093/llc/fqv046.

Hutchinson, T. 2020. "Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing." *Records Management Journal* 30 (2): 155–74. https://doi.org/10.1108/RMJ-09-2019-0055.

2001***Internet Archive: Wayback Machine***. Internet Archive. https://web.archive.org/.

Jackson, Stephen J. 2014. "Rethinking Repair." In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by T. Gillespie, P. J. Boczkowski, K. A. Foot, Cambridge, Massachusetts, USA, 221–239. The MIT Press.

Jaillant, L. 2022. "More Data, Less Process: A User-Centered Approach to Email and Born-Digital Archives." *The American Archivist* 85 (2): 533–55. https://doi.org/10.17723/2327-9702-85.2.533.

Jaillant, L., and K. Aske. 2024. "Are Users of Digital Archives Ready for the AI Era? Obstacles to the Application of Computational Research Methods and New Opportunities." *ACM Journal on Computing and Cultural Heritage* 16 (4). https://doi.org/10.1145/3631125.

Jaillant, L., O. Mitchell, E. Ewoh-Opu, and M. H. Urbaneja. 2025. "How Can We Improve the Diversity of Archival Collections with AI? Opportunities, Risks, and Solutions." *AI & SOCIETY* 40 (6): 4447–59. https://doi.org/10.1007/s00146-025-02222-z.

Jaillant, L., and A. Rees. 2023. "Applying AI to Digital Archives: Trust, Collaboration and Shared Professional Ethics." *Digital Scholarship in the Humanities* 38 (2): 571–85. https://doi.org/10.1093/llc/fqac073.

Jansen, G., and R. Marciano. 2025. "Developing Computer Vision and Machine Learning Strategies to Unlock Government-created Records." *AI & SOCIETY* 40 (6): 4513–29. https://doi.org/10.1007/s00146-025-02231-y.

Jiang, A. Q., A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M. Lachaux, P. Stock, S Subramaniam, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavirl, T. Wang, T. Lacroix, and W. El Sayed. 2024. *Mixtral of Experts*. https://arxiv.org/abs/2401.04088.

Johnston, J. A., D. A. Wallace, and R. L. Punzalan. 2022. "Messages Sent, and Received? Changing Perspectives and Policies on US Federal Email as Record and the Limits of Archival Accountability." In *Archives in a Changing Climate-Part I & Part II*, edited by V. Frings-Hessami and F. Foscarini, Cham, Switzerland, 101–21. Springer.

Kouzelis, L. R., and O. Spantidi. 2024. "Enhancing Historical Extended Reality Experiences: Prompt Engineering Strategies for AI-Generated Dialogue." *Applied Sciences* 14 (1515): 6405. https://doi.org/10.3390/app14156405.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

Lee, B. C. G. 2018a. "Machine Learning, Template Matching, and the International Tracing Service Digital Archive: Automating the Retrieval of Death Certificate Reference cards from 40 Million Document Scans." *Digital Scholarship in the Humanities* 34 (3): 513–35. https://doi.org/10.1093/llc/fqy063.

Lee, C. A. 2018. "Computer-Assisted Appraisal and Selection of Archival Materials." *In 2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, December, 2721–24. https://doi.org/10.1109/BigData.2018.8622267.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 2020. "Retrieval-augmented Generation for Knowledge-intensive NLP Tasks." In *Proceedings of the 34th International Conference on Neural Information Processing Systems* NIPS'20, Red Hook, NY, USA.

Library of Congress. 2024. MARC. https://www.loc.gov/marc/bibliographic/.

Library of Congress. 2025. Subject Headings. https://www.loc.gov/aba/cataloging/subject/.

Liu, Y., C. Heitman, L.-K. Soh, and P. Whiteley. 2025. "Machine Learning Methods for Isolating Indigenous Language Catalog Descriptions." *AI & SOCIETY* 40 (6): 4461–71. https://doi.org/10.1007/s00146-025-02223-y.

Makhortykh, M., E. M. Zucker, D. J. Simon, D. Bultmann, and R. Ulloa. 2023. "Shall Androids Dream of Genocides? How Generative AI Can Change the Future of Memorialization of Mass Atrocities." *Discover Artificial Intelligence* 3 (1): 28. https://doi.org/10.1007/s44163-023-00072-6.

Mannheimer, S., N. Bond, S. W. H. Young, H. S. Kettler, A. Marcus, S. K. Slipher, J. A. Clark, Y. Shorish, D. Rossmann, B. Sheehey. 2024. "Responsible AI Practice in Libraries and Archives: A Review of the Literature." *Information Technology and Libraries* 43 (3). https://doi.org/10.5860/ital.v43i3.17245.

Marcus, M. 1993. *Building a Large Annotated Corpus of English: The Penn Treebank*, Fort Belfort, Virginia, USA. Defense Technical Information Center. https://doi.org/10.21236/ADA273556.

Martínez-Cardama, S., and A. R. Pacios. 2022. "National Archives' Priorities: An International Overview." *Archival Science* 22 (1): 1–42. https://doi.org/10.1007/s10502-021-09367-y.

McDonald, G. 2019. "A Framework for Technology-Assisted Sensitivity Review: Using Sensitivity Classification to Prioritise Documents for Review." PhD, University of Glasgow. https://doi.org/10.5525/gla.thesis.41076.

McDonald, G., C. Macdonald, I. Ounis, and T. Gollins. 2014. "Towards a Classifier for Digital Sensitivity Review." In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval*, Vol. 8416, 500–06. Berlin, Heidelberg: ECIR. https://doi.org/10.1007/978-3-319-06028-6_48.

McKemmish, S., M. Piggott, B. Reed, and F. Upward. 2005. *Archives: Recordkeeping in Society, Amsterdam, Netherlands*. Elsevier Science & Technology. http://ebookcentral.proquest.com/lib/umichigan/detail.action?docID=1640033.

The Medici Archive Project. 2025. *Mission*. https://www.medici.org/about/mission/.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." ArXiv:1301.3781 [cs].. https://doi.org/10.48550/arXiv.1301.3781

Millar, L. A. 2017. "*The Uses of Archives*." In Archives: Principles and Practices, edited by G. Yeo, London, England, UK, 67–78. Facet.

Mitchell, T. M. 1997. *Machine Learning*, Columbus, Ohio, USA. McGraw-Hill Series in Computer Science. McGraw-Hill.

Mordell, D. 2019. "Critical Questions for Archives as (Big) Data." *Archivaria* 87 (87): 140–61.

Moss, M., and T. Gollins. 2017. "Our Digital Legacy: An Archival Perspective." *Journal of Contemporary Archival Studies* 4 (2). https://elischolar.library.yale.edu/jcas/vol4/iss2/3.

Muldoon, J., C. Cant, W. Boxi, and M. Graham. 2024. "A Typology of Artificial Intelligence Data Work." *Big Data & Society* 11 (1): 20539517241232632. https://doi.org/10.1177/20539517241232632.

Muresan, G., and D. J. Harper. 2004. "Topic Modeling for Mediated Access to Very Large Document Collections." *Journal of the American Society for Information Science and Technology* 55 (10): 892–910. https://doi.org/10.1002/asi.20034.

Nair, A., Z. R. Goh, T. Liu, and A. Y. Huang. 2025. "Web Archives Metadata Generation with GPT-4o: Challenges and Insights." *Information Technology and Libraries* 44 (2). https://doi.org/10.5860/ital.v44i2.17305.

National Archives. 2016. https://www.archives.gov/publications/general-info-leaflets/1-about-archives.html.

Nelson, L. K. 2022. "Situated Knowledges and Partial Perspectives: A Framework for Radical Objectivity in Computational Social Science and Computational Humanities." *New Literary History* 54 (1): 853–77. https://doi.org/10.1353/nlh.2022.a898331.

Nesmith, T. 2002. "Seeing Archives: Postmodernism and the Changing Intellectual Place of Archives." *The American Archivist* 65 (1): 24–41. https://doi.org/10.17723/aarc.65.1.rr48450509r0712u.

Nguyen, H. D., T.-H. A. Nguyen, and T. B. Nguyen. 2025. "A Proposed Large Language Model-Based Smart Search for Archive System." In *Information and Communication Technology*, edited by W. Buntine, M. Fjeld, T. Tran, M.-T. Tran, B. H. T. Thanh, and T. Miyoshi, Singapore, 210–23. Springer Nature. https://doi.org/10.1007/978-981-96-4285-4_18.

Nielsen, I. E., D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya. 2022. "Robust Explainability: A Tutorial on Gradient-based Attribution Methods for Deep Neural Networks." *IEEE Signal Processing Magazine* 39 (4): 73–84. https://doi.org/10.1109/msp.2022.3142719.

Nissenbaum, H. 2011. "A Contextual Approach to Privacy Online." *Daedalus* 140 (4): 32–48. https://doi.org/10.1162/DAED_a_00113.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

Ogden, J., S. Halford, and L. Carr. 2017. "Observing Web Archives: The Case for an Ethnographic Study of Web Archiving." In *Proceedings of the 2017 ACM on Web Science Conference*, June, Troy, New York, USA, 299–308. https://doi.org/10.1145/3091478.3091506.

Ogden, J., and E. Maemura. 2021. "'Go Fish': Conceptualising the Challenges of Engaging National Web Archives for Digital Research." *International Journal of Digital Humanities* 2 (1): 43–63. https://doi.org/10.1007/s42803-021-00032-5.

Osterberg, G. 2017. *Update on the Twitter Archive at the Library of Congress*. Library of Congress. https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf.

Ouyang, L., W. Jeffrey, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Siemens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. 2022. "Training Language Models to Follow Instructions with Human Feedback." In *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Vol. 35. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Poole, A. H., and A. Todd-Diaz. 2022. "'I'm Not a Very Good Visionary': Challenge and Change in Twenty-first Century North American Archival Education." *Archival Science* 22 (4): 585–616. https://doi.org/10.1007/s10502-022-09389-0.

Potter, A., and C. Saccucci. 2024. *Toward Piloting Computational Description*. https://labs.loc.gov/static/labs/work/experiments/Documents/ECD2_Final_Presentation.pdf.

Prescott, A., and A. Wiggins. 2023. *Archives: Power, Truth, and Fiction*, Oxford, England, UK. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198829324.001.0001.

Radford, A., W. Jeffrey, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* 1 (8): 9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21 (140): 1–67. https://doi.org/10.1214/10-BA521.

Raji, D., E. Denton, E. M. Bender, A. Hanna, and A. Paullada. 2021. "AI and the Everything in the Whole Wide World Benchmark." In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, edited by J. Vanschoren, and S. Yeung, Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Ranade, S. 2016. "Traces Through Time: A Probabilistic Approach to Connected Archival Data." In *2016 IEEE International Conference on Big Data (Big Data)*, 3260–65. https://doi.org/10.1109/BigData.2016.7840983.

Ranade, S. 2021. *Using AI for Digital Records Selection in Government*. https://cdn.nationalarchives.gov.uk/documents/using-ai-digital-selection-in-government.pdf.

Rasmussen, C. H. 2019. "Is Digitalization the Only Driver of Convergence? Theorizing Relations between Libraries, Archives, and Museums." *Journal of Documentation* 75 (6): 1258–73. https://doi.org/10.1108/JD-02-2019-0025.

Reusens, M., A. Adams, and B. Baesens. 2025. "Large Language Models to Make Museum Archive Collections More Accessible." *AI & SOCIETY* 40 (6): 4485–97. https://doi.org/10.1007/s00146-025-02227-8.

Risi, J., A. Sharma, R. Shah, M. Connelly, and D. J. Watts. 2019. "Predicting History." *Nature Human Behaviour* 3 (9): 906–12. https://doi.org/10.1038/s41562-019-0620-8.

Rolan, G., G. Humphries, L. Jeffrey, E. Samaras, T. Antsoupova, and K. Stuart. 2019. "More Human than Human? Artificial Intelligence in the Archive." *Archives and Manuscripts* 47 (2): 179–203. https://doi.org/10.1080/01576895.2018.1502088.

Schmidt, L. 2011. "Preserving the H-Net Email Lists: A Case Study in Trusted Digital Repository Assessment." *The American Archivist* 74 (1): 257–96. https://doi.org/10.17723/aarc.74.1.u2jw67r7257wqw66.

Schneider, J., C. Adams, S. DeBauche, R. Echols, C. McKean, J. Moran, D. Waugh. 2019. "Appraising, Processing, and Providing Access to Email in Contemporary Literary Archives." *Archives & Manuscripts* 47 (33): 305–26. https://doi.org/10.1080/01576895.2019.1622138.

Schneider, J., and P. Chan. 2016. "Let the Entities Describe Themselves." https://saaers.wordpress.com/2016/05/03/let-the-entities-describe-themselves/.

Schoenebeck, S., and P. Conway. 2020. "Data and Power: Archival Appraisal Theory as a Framework for Data Preservation." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2): 1–18. https://doi.org/10.1145/3415233.

Schwartz, J. M., and T. Cook. 2002. "Archives, Records, and Power: The Making of Modern Memory." *Archival Science* 2 (1): 1–19. https://doi.org/10.1007/BF02435628.

Shabou, B. M., J. Tièche, J. Knafou, and A. Gaudinat. 2020. "Algorithmic Methods to Explore the Automation of the Appraisal of Structured and Unstructured Digital Data." *Records Management Journal* 30 (2): 175–200. https://doi.org/10.1108/RMJ-09-2019-0049.

Shah, C., and E. M. Bender. 2022. "Situating Search." In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, CHIIR'22, New York, NY, USA, 221–32. https://doi.org/10.1145/3498366.3505816.

Shilling, C. G. 2006. "Electronic Discovery: Litigation Crashes into the Digital Age." *The Labor Lawyer* 22 (2): 207–32.

Soni, N., H. Schwartz, J. Sedoc, and N. Balasubramanian. 2024. "Large Human Language Models: A Need and the Challenges." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational*

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.

Linguistics: Human Language Technologies (Volume 1: Long Papers), edited by K. Duh, H. Gomez, and S. Bethard, Mexico City, Mexico, 8631–46. Association for Computational Linguistics. https://aclanthology.org/2024.naacl-long.477

Stephanie, B., A. Fulmer, R. Oliveira-abbey, and M. R. Rinn. 2024. "'The Career Does Not Love You Back': Impacts of Contingent Employment on Workers, Cultural Heritage Institutions, and the Archival Profession." *The American Archivist* 87 (1): 131–54. https://doi.org/10.17723/2327-9702-87.1.131.

Strubell, E., A. Ganesh, and A. McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by A. Korhonen, D. Traum, and L. Màrquez, Florence, Italy, 3645–50. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355.

Summers, E., and R. Punzalan. 2017. "Bots, Seeds and People: Web Archives as Infrastructure." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, February, Portland, Oregon, USA, 821–34. https://doi.org/10.1145/2998181.2998345.

Thibodeau, K. 2009. "The Survival of Records (And Records Management) in the Twenty-first Century." In *Proceedings of DigCCurr2009 Digital Curation: Practice, Promise and Prospects*, 26.

Toth, G. M. 2021. "Recovering and Rendering Silenced Experiences of Genocides: Testimonial Fragments of the Holocaust." *Digital Scholarship in the Humanities* 36 (1): i124–36. https://doi.org/10.1093/llc/fqaa025.

Toth, G. M., R. Albrecht, and C. Pruski. 2025. "Explainable AI, LLM, and Digitized Archival Cultural Heritage: A Case Study of the Grand Ducal Archive of the Medici." *AI & SOCIETY* 40 (6): 4561–73. https://doi.org/10.1007/s00146-025-02238-5.

Trace, C. B. 2021. "Maintaining Records in Context? Disrupting the Theory and Practice of Archival Classification and Arrangement." *The American Archivist* 83 (2): 322–72. https://doi.org/10.17723/0360-9081-83.2.322.

Trace, C. B. 2022. "Archival Infrastructure and the Information Backlog." *Archival Science* 22 (1): 75–93. https://doi.org/10.1007/s10502-021-09368-x.

Tudico, A. 2022. "National Archives Tops 200 Million Digitized Pages in Online Catalog." *National Archives*. https://www.archives.gov/news/articles/catalog-200-million-digitized-pages.

Turner, A. M., E. D. Liddy, and J. Bradley. 2002. "Breathing Life into Digital Archives: Use of Natural Language Processing to Revitalize the Grey Literature of Public Health." In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL'02, July, New York, NY, USA, 411. https://doi.org/10.1145/544220.544359.

Underwood, T. 2014. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127 (1): 64–72. https://doi.org/10.1525/rep.2014.127.1.64.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, et al., Vol. 30., Long Beach, CA, USA, 1–11. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vellino, A., and I. Alberts. 2016. "Assisting the Appraisal of E-mail Records with Automatic Classification." *Records Management Journal* 26 (3): 293–313. https://doi.org/10.1108/RMJ-02-2016-0006.

Venkata, S. K., P. Young, M. Bell, and A. Green. 2022. "Alexa, Is This a Historical Record?" *Journal on Computing and Cultural Heritage* 15 (1): 1–20. https://doi.org/10.1145/3479008.

Verhagen, M., K. Lynch, K. Rim, and J. Pustejovsky. 2022. "The CLAMS Platform at Work: Processing Audiovisual Data from the American Archive of Public Broadcasting." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, 2498–506. European Language Resources Association. https://aclanthology.org/2022.lrec-1.266.

Wang, A., J. Morgenstern, and J. P. Dickerson. 2024. *Large Language Models Cannot Replace Human Participants Because They Cannot Portray Identity Groups*. https://arxiv.org/abs/2402.01908.

Wiedeman, G. 2019. "The Historical Hazards of Finding Aids." *The American Archivist* 82 (2): 381–420. https://doi.org/10.17723/aarc-82-02-20.

Windon, K., and J. Youngblood. 2024. "Privacy Considerations in Archival Practice and Research." In *Human Privacy in Virtual and Physical Worlds: Multidisciplinary Perspectives*, edited by M. C. Lacity, and L. Coon, Cham, Switzerland, 205–34. Technology, Work and Globalization. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-51063-2.

Wright, H., T. N. L. Evans, and K. Green. 2023. "NLP and Archaeology: A View from A Digital Archive." In *Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches.*, edited by C. Gonzalez-Perez, P. Martin-Rondilla, M. Pereira-Fraiña, New York City, New York, USA, 215–28. Quantitative Archaeology and Archaeological Modelling. Springer International Publishing. https://doi.org/10.1007/978-3-031-37156-1_10.

Yakel, E. 2003. "Archival Representation." *Archival Science* 3 (1): 1–25. https://doi.org/10.1007/BF02438926.

Yeo, G. 2012. "The Conceptual Fonds and the Physical Collection." *Archivaria*, April, 43–80.

Yoxall, H. 1984. "Privacy and Personal Papers." *Archives & Manuscripts* 12, 30–44.

Zhang, S., S. Peng, P. Wang, and J. Hou. 2024. "Archives Meet GPT: A Pilot Study on Enhancing Archival Workflows with Large Language Models." In *iConference 2024 Proceedings*, March. https://hdl.handle.net/2142/122806.

Cite this article: Dunagan, L., Johnston, J. and Card, D. 2026. Missing context: a review of natural language processing for archives. *Cambridge Forum on AI: Culture and Society*, **2**, e25, 1–21. https://doi.org/10.1017/cfc.2026.10031

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 16:00:18, subject to the Cambridge Core terms of use.