InterPARES Trust AI

Teachable AI for the Archival Professions – Module 3:

AI/ML for processing textual records in Archives

Kaila Fewster[1] and Richard Arias-Hernandez[2]

This educational module is part of a series of learning materials developed by InterPARES Trust AI[3] researchers and educators to train archival professionals and students to effectively leverage artificial intelligence in their archival work. The current version was completed on April 11th, 2025.

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

[2] Associate Professor of Teaching, School of Information, University of British Columbia and InterPARES AI Trust Researcher. richard.arias@ubc.ca

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

[4] Teachable AI for the Archival Professions – Module 3: AI/ML for processing textual records in Archives © 2025 by Fewster, Kaila and Arias-Hernandez, Richard is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/

Module 3: AI/ML for processing textual records in Archives

READ FOR THIS MODULE

Required:

- Hutchinson, T. (2020). Natural language processing and machine learning as practical toolsets for archival processing. *Records Management Journal*, *30*(2), 155–174. https://doi.org/10.1108/RMJ-09-2019-0055

- Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2024).Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, *56*(2), 1–47. https://doi.org/10.1145/3604931

- Muehlberger, G., Seaward, L., Terras, M., Oliveira, S. A., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G.,Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., … Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, *75*(5), 954–976. https://doi.org/10.1108/JD-07-2018-0114

Recommended:

- Allen, M. (1987). Optical Character Recognition: Technology with New Relevance for Archival Automation Projects. *The American Archivist*, *50*(1), 88–99. https://www.jstor.org/stable/40294351

- Clough, P., Tang, J., Hall, M. M., & Warner, A. (2011). Linking archival data to location: A case study at the UK National Archives. *Aslib Proceedings*, *63*(2/3), 127–147. https://doi.org/10.1108/00012531111135628

- Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage*, *15*(1), 4:1-4:15. https://doi.org/10.1145/3479010

- Marciano, R. (2022). Afterword: Towards a new Discipline of Computational Archival Science (CAS). In L. Jaillant (Ed.), *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections* (1st ed., Vol. 2, pp. 205–216). Bielefeld University Press. https://doi.org/10.14361/9783839455845

- Mordell, D. (2019). Critical Questions for Archives as (Big) Data. *Archivaria*, *87*, 140–161. https://archivaria.ca/index.php/archivaria/article/view/13673

OVERVIEW

This module provides an overview of how artificial intelligence (AI) and machine learning (ML) tools have been and continue to be used in archives for processing textual records and illustrates how archival professionals can critically engage with these tools. In particular, this module discusses the origins of AI-based textual record processing in the archive with Optical Character Recognition (OCR), Handwritten Text Recognition (HTR), and computer vision. Moreover, it also presents Natural Language Processing (NLP), Named Entity Recognition (NER), and topic modelling as more advanced applications of AI for textual record processing, which are all growing in popularity. Finally, it highlights some of the challenges associated with integrating

AI tools into textual record processing, and advocates for consistent and active human oversight.

LEARNING OBJECTIVES

By the end of this lesson, students will be able to:

- Explain the history of AI tools in archives with emphasis on optical character recognition (OCR), handwritten text recognition (HTR), digitization, and computer vision.
- Use AI tools for the treatment of text-based digital assets in the archives including digitized analogue records and born-digital records.
- Identify opportunities to use AI/ML tools for archival processing of digital text-based records.

Introduction

With origins in the mid-20th century, artificial intelligence (AI) and machine learning (ML) tools have been a part of archival workflows for a long time, albeit not in the same way these tools are implemented into modern workflows today. For instance, AI-based tools like computer vision and optical character recognition (OCR) have both been used in archives since the early 1980s and have become essential processes in the digitization of analogue archival records (Allen, 1987; Pugh, 2024). In this sense, it is relevant for information professionals to know how these tools have historically been used in the archive to feel more confident using contemporary versions themselves and to be better prepared for the future advancements of these technologies.

Similarly, working with AI tools in the archive requires a reconceptualization of the records as data, in the sense that the archive serves as data for the AI tools' algorithms. Thinking about archives in this way is necessary as digitalization, which refers to the increased adoption of digital technologies and the subsequent effects on society, continues worldwide (Sengsavang, 2023). Digitization of analogue materials for access and/or preservation along with the surge in born-digital records represents the ongoing digitalization of the archive and growing reliance in these spaces on digital technologies, like AI and ML, for the processing, preservation, and access to these records. In this context, there are several archival processes for textual records (e.g., manuscripts, correspondence, documents, emails, etc.), such as: transcription, full-text indexing, and automated metadata extraction, which are increasingly possible through AI tools like OCR and Handwritten Text Recognition (HTR). Therefore, archivists and records managers interested in implementing AI into their workflows should first evaluate whether some of the existing systems they use are already powered by contemporary AI, or whether they could be enriched through the integrations of these tools.

However, it is also worth noting that these technologies are not perfect and can be prone to mistakes if not properly trained or applied. Furthermore, like any algorithmic system, tools like computer vision are prone to social and cultural biases, which must be identified and mitigated by the human operators of these technologies. In this sense, while AI/ML tools have major potential in improving archival workflows for processing textual records, archivists and records managers must still be actively and continuously involved in the process to ensure accurate and unbiased outcomes.

History of AI/Computer Vision for processing textual records in Archives

As mentioned in the introduction, AI tools like Optical Character Recognition (OCR) and computer vision have been used in archival settings since the mid 1980s, most notably in digitization workflows (Allen, 1987). Based on several experiments with OCR technology at the United States National Archives and Records Administration (NARA), a report published in 1983 notes that OCR has significant potential for improving the availability, usefulness, and cross-references options of printed finding aids (Allen, 1987). This report ultimately inspired several further pilot projects investigating OCR in archives, including one undertaken by NARA to test early iterations of Handwritten-Text Recognition (HTR) in 1986 (Allen, 1987).

In general, Optical Character Recognition (OCR) can be understood as the automated process of extracting information, usually text, from scanned documents and converting it into machine-readable form (Wang et al., 2021). It is mostly applied to typed/printed documents and its models rely on character recognition based on specific fonts and languages. The process requires the software to recognize the optical characters in the image first, before it translates the extracted images into digital representations of text (e.g. ASCII or Unicode) that can be searched and indexed (Wang et al., 2021). Early OCR translated each character individually, which often led to recognition errors; however, modern models use tools like convolutional neural networks that extract meaningful machine-readable features from the image, like patterns, lines, and shapes, which are then translated back into text by a text-generating model (Wang et al., 2021). Furthermore, OCR serves as the foundation for most archival digitization workflows by making documents searchable, and more widely available for use online (Leviner, 2023). While archivists and records managers do not necessarily need to know exactly how modern OCR works, it is still important to understand its extensive history in the archive.

A particularly important type of OCR for archives is HTR, which can generate machine-readable text from scanned images of handwriting. With roots in early OCR-based archival projects, like those taken on by NARA in the 1980s, HTR tools have improved considerably over the last 35 years, and have become much more stable, accurate, and efficient (Terras, 2022). As many archival documents are manuscripts, HTR makes it possible for significantly more records to be digitized, indexed, searchable, and made available online.

Likewise, computer vision refers to the ability of computers to identify, extract and understand objects in digital images or video (IBM, 2021). Taking that into consideration, OCR and HTR fall under the broader umbrella of computer vision, as it relies on AI models to extract text from images and convert it to machine-readable form. As such, OCR's inception in the 1970s and introduction into the archives in the 1980s represents computer vision's first practical application in archival contexts and more broadly (IBM, 2021). Since then, the focus of computer vision study has turned to object recognition, which will be further discussed in another module (i.e., AI/ML techniques for processing image records). Nonetheless, as the fundamental technology behind OCR, computer vision serves a valuable role in the process of archival digitization.

Anyone familiar with archival records knows how valuable they can be for research, study and learning. However, these documents can often become fragile over time, due to use or often physical and chemical instability in the materials themselves. As such, with the rise of institutional use of OCR in the 1980s, digitization became an essential part of the archival workflow not only to replicate documents to preserve them digitally, but also to provide wider access to materials that otherwise would require an in-person visit (Terras, 2011). Digitization is the process of converting analogue materials to digital format, which in an archival setting, is typically completed using scanners, OCR, and other computer vision software (Behler, 2022). As mentioned earlier, NARA was an early archival adopter of OCR and digitization, but they were not alone. In 1986, the Archivo General de Indias, a 200-year-old archive in Spain, with materials documenting the Spanish Empire in the Philippines and Americas, undertook a massive project to digitize the archives holdings (Terras, 2011). By 1992 the archive had digitized over 7 million pages, and the project reached 11 million pages digitized by 1998 (Terras, 2011). From there, and with the introduction of the internet, digitization projects became increasingly popular, and the concepts of digital libraries and collections became more mainstream (Terras, 2011). In this sense, as AI technologies continue to improve, so too does the digitization process in the archive, as their technological advances are interrelated.

TEST YOUR KNOWLEDGE

1. How has the introduction of OCR and HTR in the archive influenced the accessibility of archival records?
2. What are some of the potential benefits and drawbacks to digitization in archives?
3. Although OCR and computer vision technologies have vastly improved over the last 35 years, what types of archival documents may still be difficult to process with these models? Why? Are recent developments in AI changing this situation?

Archives as Data

With the history of these processes in mind, it’s also worth acknowledging that working with AI tools for digitization or other archival processes requires a reconceptualization of the documents not just as records, but also as data.

This reconceptualization is discussed in computational archival science (CAS), a proposed transdiscipline field of study "concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access…" (Marciano et al., 2018). CAS relies on a blend of archival thinking and computational methods, which encompasses using AI tools like OCR, HTR, and computer vision. Moreover, scholars in the field suggest that harnessing data science advancements is useful in combating the 'dark archives', records that are saved for future use, but only accessible by the custodian, held in many institutions. (Marciano, 2018). In this sense, digitization has long been a form of CAS, as creating digital surrogates to be processed, analyzed and used constitutes a form of datafication of archival documents. Nonetheless, it's important to remember, as Mordell suggests, that "digital archives are not already data by virtue of being digital but *become* data – are datafied – through the various acts of preparing them for manipulation by computational means" (Mordell, 2019). In other words, archival documents become data once they are manipulated through computational methods, thus contributing to a new digital outcome. Following this train of thought, Cordell even argues that OCR'd and digitized materials should be considered new versions of their source texts as they present not only unique insights into the document portrayed, but also into the creation of the digital copy itself (Cordell, 2017). In this sense, digitized records from archives are not records themselves but instead copies of records and, in some cases, augmented copies that include features not readily available to human readers from the original records. The analog document still maintains the features of authenticity and reliability that make the record valuable. Therefore, the digitized copy primarily exists for accessibility and not necessarily to preserve the integrity of the record (Duranti et al., 2022).

Notwithstanding the concerns around digital authenticity, understanding records processed by AI tools as data is also increasingly relevant as digitalization, as set forth in the introduction, becomes commonplace in business and society. As digital environments continue to grow, archives are required to adapt by using old AI technologies and new ones to best preserve records and make them accessible. In this sense, automating archival processes like transcription, description, and preservation is helpful in improving archivist and records managers' workloads and encouraging the use of more computational methods in the field. However, it is also important to recognize much of our cultural heritage will likely never be digitized, and thus a high-quality digital archive lies not only in digitization, but also in access to high-quality descriptive metadata and finding aids (Zaagsma, 2023).

As has been discussed so far throughout the module, digitization is an essential archival process that has long made use of AI tools and is inextricably linked to these tools' development in the archive. In particular, OCR and HTR technologies have been at the forefront of archival experiments with AI because of their applicability to the large swaths of textual records held in these collections. As a result of these advancements, digitization of archival records serves a myriad of purposes, including improving accessibility, arrangement and description, and preservation initiatives. First of all, digitizing records allows for asynchronous use, making archival research more feasible for those unable to access the records in-person (Kenely et al., 2016). Furthermore, the ability to perform full-text searches on OCR'd records (Kenely et al., 2016) and index OCR'd finding aids (Allen, 1987) help users find highly specific information unique to their needs. Second, digitizing archival documents can also be beneficial for improving descriptions as the OCR'd text can be included directly in the finding aid or be fed to a LLM to generate summaries. Additionally, digitized documents can be further ingested into other AI models to augment descriptions and provide third-order access to digitized fonds (Lemieux, 2014), as was done with the Cybernetic Thought Collective's project investigating computational archival methods of representing community provenance across fonds (Anderson, 2021). In this sense, digitizing materials serves as a jumping off point for further computational manipulation of archival documents. Finally, digitization has long been used as a tool in archives for digital preservation. It should be noted here that there is a difference between digitization and digital preservation. Digitization is a conversion process, where analog materials are converted to a digital format. On the other hand, digital preservation is an active and ongoing process of safeguarding digital data (digitized data or otherwise) from corruption, deletion, or obsolescence for extended, if not indefinite, periods of time (Behler, 2022). As such, should archivists be interested in digital preservation for analog records, digitization serves as a foundational process for subsequent preservation steps, like ingestion into a secure digital repository.

Born-digital records, even though they do not benefit from OCR, can benefit from other AI technologies, and are understood as records that have been natively created in a digital format (The National Archives, n.d.). In recent years, an increasing number of case studies and applications have shown the potential benefit of using AI tools in the archives (Lee; 2018; Ranade, 2016). As business is increasingly conducted digitally, subsequent records like emails, datasets, databases, code for algorithms, spreadsheets, and PDFs still must be kept, organized, and indexed even if they do not tangibly exist. In this sense, AI tools can be helpful for automating processes like appraisal and metadata creation as archivists become the stewards of ever-growing digital archives (Colavizza et al., 2021). Moreover, the possibility to extract archival content from born-digital records to create aggregate datasets which can be made accessible online invites new interaction with materials from all around the world (Jaillant et al., 2022). Therefore, as archives continue to collect born-digital records, it is important for archivists and records managers to recognize how these records can be used as data, and moreover, the ways AI/ML can be used as valuable tools for processing this massive amount of data into more accessible and interpretable formats (Moss et al., 2018).

Processing Textual Records with AI/ML tools

As previously mentioned, OCR was the first type of AI integrated into archival workflows. However, not only has OCR technology improved, but AI tools more broadly have seen enormous breakthroughs in the last decade. As such, many of these tools have begun to be implemented in the archives for processing both analog and born-digital records. OCR, for instance, has evolved to recognize and transcribe many different scripts and languages, detect text on heavily degraded documents, and even reconstruct fonts from historical records (Al-Kaffaf et al., 2012; Dutta et al., 2012). Moreover, it is also increasingly being used as a tool for illustrating connected provenances among records using full-text search and indexing (Anderson, 2021; Travis et al., 2016, Paolanti et al., 2022).

ACTIVITY #1

This activity demonstrates OCR using pytesseract, open source Python Library. Students will use a pre-built Python code available on Google Colab to upload an image of typed text: OCR_Tesseract_Activity.ipynb

Students should make a copy of the code and save it to their Drive. The instructor then guides the activity showing step-by-step how to run the different sections of Google Colab. For this activity, the instructor may provide a JPG file to be OCRd for the whole class. If students are using their own JPG rastered image to be OCR'd, then the instructor needs to guide them to update the file name on the Python code section "Running Tesseract-OCR on the uploaded rastered image file." After the activity, the students discuss in groups what the software got right versus wrong and why.

Even more impressive are the advancements in HTR, a field which has seen dramatic growth in the last 25 years with the introduction of advanced pattern recognition in the 1990s and the subsequent use of neural networks in the 2000s and 2010s (Muehlberger et al. 2019). The Transkribus project, launched in 2015, is an excellent example of an ongoing HTR project built through collaboration between archivists, (digital) humanities scholars, and computer scientists. The project functions through memory institutions, scholars and even the public providing digitized images and transcripts to train the Transkribus' HTR neural network model, while computer scientists contribute to maintaining the technology through research (Muehlberger et al., 2019). Moreover, given that the software is built with ML technology, the platform gets better with every document processed, providing tangible benefits to their contributors through more accessible digital collections, and a wealth of data for further research (Muehlberger et al., 2019). Not only is Transkribus primarily free-to-use online, but most of the code is also open-source, meaning other organisations can use their models as a basis to train their own specialized HTR software (Muehlberger et al., 2019). In this sense, Transkribus demonstrates the potential of using HTR models in archives and represents an interdisciplinary approach to improving access to digitized handwritten documents.

ACTIVITY #2

To get acquainted with Transkribus, students first need to watch the 11 videos (~approx. 30min) in the "Getting Started with Transkribus" playlist on their Youtube channel (Getting Started with Transkribus - YouTube), and make a free account on their site (https://www.transkribus.org/).

Then, using the instructions from the Youtube tutorials and/or guidance of the instructor during a walkthrough, students can upload their own JPGs of manuscripts (or ones provided by the instructor - we recommend using copies of actual records from archives) and experiment with the different models Transkribus has to offer. Afterwards, students can discuss in groups what did and did not work well, and why.

Natural Language Processing for Digital-born or Digitized Records

Another popular machine learning tool growing in popularity in the archive is Natural Language Processing (NLP). NLP is a sub-field of AI that refers to a computer's ability to understand data encoded in natural language, which is any language that humans use, as opposed to computer languages, like code (Goodman, 2019). InterPARES defines NLP as "AI methods which are applied to human languages, including in spoken and written forms." (InterPARES, 2025). NLP uses regular expressions, which are defined sequences of characters used in algorithms for pattern-matching that enable computers to identify grammatical, syntactic, and semantic units within a text, and will be explored in further depth below (Hutchinson, 2020). Using NLP in archives, especially on born-digital records, presents opportunities for improving appraisal and selection workflows, identifying personal or sensitive information, as well as for improving description and access through metadata extraction (Clough et al., 2011; Hutchinson, 2020; Lee, 2018). For instance, ePADD (or Email Processing, Appraisal, Discovery, Delivery) is an open-source software developed by Stanford University's Special Collections and University Archives, and launched publicly in 2015 (Schneider et al., 2019). Tailored towards email archives, ePADD was designed to specifically address the challenge of processing large volumes of data that may have cultural or historical value (Schneider et al., 2019). Using a custom NLP toolkit designed for ePADD and connected to external datasets like OCLC and the Library of Congress Subject Headings/Name Authority Files, the software is able to identify entities within the text (Hutchinson, 2020). Moreover, ePADD is also able to perform ‘Lexicon Analysis,’ which allows for keyword searching based on regular expression patterns, themes, or entities (Schneider et al., 2020). In this sense, although primarily restricted to born-digital email archives, ePADD illustrates how NLP has significant potential for improving appraisal workflows for archivists, and making it easier to develop more detailed record descriptions.

Named Entity Recognition (NER) is another AI tool increasingly being used in archival and records management contexts. NER identifies predefined categories of objects in a body of text, including names of individuals, organizations, locations, expressions of times, etc. (IBM, 2023). NER is a subset of NLP. NLP processes natural language into a form that computers can understand, and NER is the subsequent process where named entities identified within the text are classified based on predefined categories (Ehrmann et al, 2024). Like NLP, NER has historically relied on regular expressions to match patterns within a text, but more modern tools rely on Deep Learning models which predict whether certain word sequences represent entities (Cunha and Ramalho, 2021). In this sense, over the past twenty years NER has undergone a significant evolution from simply entity recognition and classification to entity disambiguation and linking, which represents the progression of information extraction from a “document- to a semantic-centric viewpoint” (Ehrmann et al., 2024). For archivists and records managers, NER has potential for enhanced information retrieval, improved descriptions, automated extraction of metadata, assisted sensitivity analysis and redaction of copies of records, but also for visualizing connections between records and fonds (Anderson, 2021).

For instance, InterPARES researcher Basma Makhlouf Shabou and colleagues used NER as part of their study investigating the development for automated appraisal methods for structured and unstructured archival data (Makhlouf Shabou et al., 2020). Using a large amount of complex and diverse data from the State Archives of Neuchâtel, NER was performed on the dataset in order to extract dates, locations, services names, and personal names based on the 'Europeana Newspapers NER' corpora (Makhlouf Shabou et al, 2020). With this method, they were able to index over 19,000 documents and illustrate a meaningful proof of concept for using AI to make defensible automated appraisal decisions for archival retention and disposition (Makhlouf Shabou et al., 2020). Another interesting use of NER in archives come from Portugal, where Cunha and Ramalho trained and tested several different 'off-the-shelf' NER and NLP algorithms on a large corpus of archival finding aids to investigate the accuracy of these models on identifying personal names, locations, and major events (Cunha and Ramalho, 2021). They found that, when trained on similar data to those they ultimately ingested for NER, the models were able to provide relatively accurate results, illustrating how these technologies could be used within finding aid databases to improve discoverability and better visualize relationships between fonds (Cunha and Ramalho, 2021). Finally, in an attempt to address the challenges of colonial archives, including highlighting those who have previously been marginalized in the records, Luthra et al. developed a new annotated entity typology which expands upon the types of entities that can be identified within the records to include both valued and undervalued individuals, with further qualifiers about gender, legal status, and other defining attributes (Luthra et al., 2024). The typology was tested on testaments from Dutch East India Company archive, held at the National Dutch archives, using a fit-for-purpose Dutch LLM to evaluate how annotations can enable NER to highlight more diverse entities (Luthra et al., 2024). Ultimately, this work shows that although some entities were more difficult to detect than others, it is possible to expand NER capabilities to tag and characterize unnamed historical entities even in complex archival documents and reaffirms that this work is necessary for highlighting gaps within colonial archives.

ACTIVITY #3a

Watch YouTube video of ePADD version 9 demo (7:33 mins): ePADD Version 9 Demo, read the InterPARES case study here, and then discuss support for appraisal and sensitivity analysis which relies on NLP and NER.

ACTIVITY #3b

This activity demonstrates NLP and NER as supported by spaCy, an open-source Python Library. Students will use a pre-built Python code available on Google Colab to upload a plain text file, which could be the result of the previous OCR or HTR activity: Spacy.ipynb

Students should make a copy of the code and save it to their Drive. The instructor then guides the activity showing step-by-step how to run the different sections of Google Colab. For this activity, the instructor may provide the plain text (TXT) file for this activity, or students can try it with one of their own. After the activity, the students discuss with the class the functionality of spaCy for processing of textual records in archives. This activity is based on the class activity “Lab 3 Recipe Natural Language Processing and Sentiment Analysis” developed by Dr. Victoria Lemieux for her ARST 500 course at the University of British Columbia.

Topic modelling is a more complex application of ML that relies on unsupervised learning, where models are trained on unlabeled data so they develop their own methods of classification and pattern recognition (van Hooland and Coeckelbergs, 2018). As such, topic modelling has gained momentum in more recent years in the digital humanities fields for “interpret[ing] very large corpora of full-text documents” through clustering keywords extracted from the documents into overarching topics or themes (van Hooland and Coeckelbergs, 2018). In this sense, topic modelling is valuable in archives for classifying thematically similar documents, as well as making topical connections between records. Previous experiments with topic modelling in archives include Blanke and Wilson's investigation of topic modelling for identifying and classifying textual records into different 'epochs' (time periods) based on the use and characteristics of language in the documents, as well as van Hooland and Coeckelbergs' work at the European Commission archives, where extracted keywords from the records were matched to an existing thesaurus to improve access and discoverability (Blake and Wilson, 2017; van Hooland and Coeckelbergs, 2018). More recently, Grant et al. used topic modelling on a corpus of archival documents from the UK, US, and United Nations High Commissioner for Refugees (UNHCR) related to refugee cases in the 1970s to identify themes and the relationships between them through time for conducting historical policy analyses (Grant et al., 2021). With these cases in mind, there is significant potential for topic modelling in archives to increase accessibility, and its capabilities are likely to improve as unsupervised ML advances.

ACTIVITY #4

This activity demonstrates topic modelling using Latent Dirichlet Allocation (LDA) algorithms. Students will use pre-built LDA code available on Google Colab to upload a CSV file:
ARST556P_LIBR582_TopicModeling.ipynb
Students should make a copy of the code and save it to their Drive. The instructor then guides the activity showing step-by-step how to run the different sections of Google Colab. For this activity, the instructor should provide a CSV file to be ingested into the model for the whole class. One example of a dataset that could be used for this activity is the Disneyland Reviews

dataset. After the activity, the students discuss in groups how accurate they found the clustered topics to be, what the algorithm got right versus wrong, and why.

Challenges and Limitations of AI/ML for Processing Textual Records

AI/ML has massive potential for improving archival workflows through automating classification and extracting keywords to improve discoverability, but there are still challenges with implementing these technologies in archives, and limitations in the technologies themselves. AI/ML technologies naturally have biased outputs that are highly dependent on the training data they receive, which can often reflect existing social inequalities and entrench them within the system's algorithms. Moreover, this can raise ethical concerns around how sensitive information is ingested into and processed by non-human actors (Ehrmann et al, 2023). When it comes to text processing in particular, OCR and HTR are heavily influenced by the condition of the original record, the quality of the digital scan, the diversity of the characters, fonts, and punctuation in the document, epoch-specific variances, and the cultural languages in the training data (Ehrmann et al., 2023). This 'OCR noise' also affects subsequent processes like NER and NLP, with van Strien et al. illustrating that different qualities of OCR'd documents can impact the accuracy of entity recognition by 20% (van Strien et al., 2020). In this sense, OCR noise presents a significant challenge to textual records processing, as it can present in many different ways depending on the record quality, which can be extremely variable in archival contexts. Additionally, NLP and NER processes can also produce unreliable results depending on the size of the corpora being used to train the models (Silberztein, 2024). Therefore, it is necessary for archivists to consider how the models they use are being trained and ensure the training corpora used are designed to support models working with archival records.

It is also important to consider that many off-the-shelf AI tools use black-box models, which means there are little to no accountability mechanisms present for evaluating the inner-workings of the algorithms. As a result, archivists and records managers must rely on existing theoretical archival frameworks of authenticity to address some of these transparency issues (Bunn, 2020). Even so, Mordell warns that conceptualizing archival records as data is a slippery slope, as it “signifies an amenability to computation of analytical purposes” (Mordell, 2019). In this sense, it is relevant for archivists to consider what types of textual records are being ingested into their AI/ML models, and how they plan to uphold the archival principles of authenticity, integrity and trustworthiness while using these technologies. As such, while they may improve efficiency and reduce backlog, working with AI/ML technologies requires consistent and active monitoring from human actors to ensure reliable and unbiased outcomes. Nonetheless, AI’s ability to classify and enhance discoverability of textual records through keyword extraction, amongst other processes, is an opportunity to enhance user access and alleviate archivists and records managers’ ever-growing workload.

MODULE COMPREHENSION ACTIVITY

Students are presented an archival collection with a variety of text-based records that hypothetically they would be asked to process (e.g., digitized handwritten letters and notebooks, scanned corporate reports and meeting minutes, a born-digital email collection, born-digital infrastructure project files, etc.). Students are tasked with writing a short recommendation to the Archives Director (1000-1500 words) to test AI tool(s) to enhance the effectiveness of processing each group of records.

Students provide a rationale for the use of these tools and potential caveats that need to be addressed.

SUMMARY

In summary, while processing textual records with AI tools is not necessarily new to archivists, the fields of OCR, HTR and computer vision have made major strides in refining these technologies over the last 20 years. As the world moves towards digitalization, archives are no exception, and more than ever archivists and records managers need tools to manage both analogue and born-digital textual records. In this sense, technologies like OCR and HTR convert textual records into machine readable formats for easier indexing and searching. Similarly, NLP serves as a way for computers to understand and process natural language and NER can identify and classify key entities, both of which can be applied to machine readable records to aid in appraisal, improving descriptions, automating metadata extraction and generation, and visualizing connections between fonds. Moreover, topic modelling is useful for classifying thematically similar documents, as it interprets data through identifying and clustering keywords and themes. Still, working with AI tools requires human management and oversight to ensure high-quality and unbiased outcomes from the models. Although there is still much progress to be made, projects like Transkribus and ePADD ultimately illustrate the massive potential of integrating AI/ML technologies into processing textual records in the archive.

REFERENCES

Allen, M. (1987). Optical Character Recognition: Technology with New Relevance for Archival Automation Projects. *The American Archivist*, *50*(1), 88–99. https://www.jstor.org/stable/40294351

Al-Khaffaf, H. S. M., Shafait, F., Cutter, M. P., & Breuel, T. M. (2012). On the performance of Decapod's digital font reconstruction. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 649–652. https://ieeexplore.ieee.org/document/6460218

Anderson, B. G. (2021). On Constructing a Scientific Archives Network Exploring Computational Approaches to the Cybernetics Thought Collective. *Archivaria*, *91*, 104–147. https://www.proquest.com/docview/2539879519/abstract/B14682379B2C47A6PQ/1

Behler, A. (2022, September). *Digitization vs. Digital Preservation* (History Associates Incorporated, Interviewer) [Interview]. https://preservica.com/resources/blogs-and-news/digitization-vs-digital-preservation

Blanke, T., & Wilson, J. (2017). Identifying epochs in text archives. *2017 IEEE International Conference on Big Data (Big Data)*, 2219–2224. https://doi.org/10.1109/BigData.2017.8258172

Clough, P., Tang, J., Hall, M. M., & Warner, A. (2011). Linking archival data to location: A case study at the UK National Archives. *Aslib Proceedings*, *63*(2/3), 127–147. https://doi.org/10.1108/00012531111135628

Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage*, *15*(1), 4:1-4:15. https://doi.org/10.1145/3479010

Cordell, R. (2017). "Q i-jtb the Raven": Taking Dirty OCR Seriously. *Book History*, *20*(1), 188–225. https://doi.org/10.1353/bh.2017.0006

Cunha, L. F. da C., & Ramalho, J. C. (2021). *NER in Archival Finding Aids*. 1–16. https://repositorium.sdum.uminho.pt/bitstream/1822/73504/2/SLATE_2021_NER_in_Archival_Finding_Aids.pdf

Duranti, L., Rogers, C., & Thibodeau, K. (2022). Authenticity. *Archives and Records*, *43*(2), 188–203. https://doi.org/10.1080/23257962.2022.2054406

Dutta, S., Sankaran, N., Sankar, K. P., & Jawahar, C. V. (2012). Robust Recognition of Degraded Documents Using Character N-Grams. *2012 10th IAPR International Workshop on Document Analysis Systems*, 130–134. https://doi.org/10.1109/DAS.2012.76

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2024). Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, *56*(2), 1–47. https://doi.org/10.1145/3604931

Goodman, M. (2019). *"What is on this disk?" An Exploration of Natural Language Processing in Archival Appraisal* [Masters Thesis, University of North Carolina at Chapel Hill]. https://cdr.lib.unc.edu/concern/masters_papers/8c97kw01h

Grant, P., Sebastian, R., Allassonnière-Tang, M., & Cosemans, S. (2021). Topic modelling on archive documents from the 1970s: Global policies on refugees. *Digital Scholarship in the Humanities*, *36*(4), 886–904. https://doi.org/10.1093/llc/fqab018

Hutchinson, T. (2020). Natural language processing and machine learning as practical toolsets for archival processing. *Records Management Journal*, *30*(2), 155–174. https://doi.org/10.1108/RMJ-09-2019-0055

IBM. (2023, August 26). *What Is Named Entity Recognition*. https://www.ibm.com/topics/named-entity-recognition

Jaillant, L., Aske, K., Goudarouli, E., & Kitcher, N. (2022). Introduction: Challenges and prospects of born-digital and digitized archives in the digital humanities. *Archival Science*, *22*(3), 285–291. https://doi.org/10.1007/s10502-022-09396-1

Kenely, M., Potter, B., West, B., Cobbin, P., & Chang, S. (2016). Digitizing Archival Records: Benefits and Challenges for a Large Professional Accounting Association. *Archivaria*, 75–100. https://archivaria.ca/index.php/archivaria/article/view/13559

Lemieux, V. L. (2014). Toward a "Third Order" Archival Interface: Research Notes on Some Theoretical and Practical Implications of Visual Explorations in the Canadian Context of Financial Electronic Records. *Archivaria*, 53–93. https://archivaria.ca/index.php/archivaria/article/view/13721

Leviner, S. (2023, July 11). The Role of OCR in Digitizing Historical and Archival Documents. *CharacTell*. https://www.charactell.com/resources/the-role-of-ocr-in-digitizing-historical-and-archival-documents/

Luthra, M., Todorov, K., Jeurgens, C., & Colavizza, G. (2024). Unsilencing colonial archives via automated entity recognition. *Journal of Documentation*, *80*(5), 1080–1105. https://doi.org/10.1108/JD-02-2022-0038

Makhlouf Shabou, B., Tièche, J., Knafou, J., & Gaudinat, A. (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

Marciano, R. (2022). Afterword: Towards a new Discipline of Computational Archival Science (CAS). In L. Jaillant (Ed.), *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections* (1st ed., Vol. 2, pp. 205–216). Bielefeld University Press / transcript Verlag. https://doi.org/10.14361/9783839455845

Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M., & Conrad, M. (2018). Archival Records and Training in the Age of Big Data. In J. Percell, L. C. Sarin, P. T. Jaeger, & J. C. Bertot (Eds.), *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education* (Vol. 44B, pp. 179–199). Emerald Publishing Limited. https://ai-collaboratory.net/wp-content/uploads/2020/10/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf

Muehlberger, G., Seaward, L., Terras, M., Oliveira, S. A., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., … Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, *75*(5), 954–976. https://doi.org/10.1108/JD-07-2018-0114

Mordell, D. (2019). Critical Questions for Archives as (Big) Data. *Archivaria*, *87*, 140–161. https://archivaria.ca/index.php/archivaria/article/view/13673

Moss, M., Thomas, D., & Gollins, T. (2018). The reconfiguration of the archive as data to be mined. *Archivaria, 86*, 118-151. Retrieved from https://www.proquest.com/scholarly-journals/reconfiguration-archive-as-data-be-mined/docview/2518872186/se-2

Ranade, S. (2016). Traces through time: A probabilistic approach to connected archival data. *2016 IEEE International Conference on Big Data (Big Data)*, 3260–3265. https://doi.org/10.1109/BigData.2016.7840983

Schneider, J., Adams, C., DeBauche, S., Echols, R., McKean, C., Moran, J., & Waugh, D. (2019). Appraising, processing, and providing access to email in contemporary literary archives. *Archives and Manuscripts*, *47*(3), 305–326. https://doi.org/10.1080/01576895.2019.1622138

Sengsavang, E. (2023, February 20). *AI-Assisted Digitization of Documentary Heritage Materials*. 2nd Symposium, ITrust AI, Abu Dhabi, United Arab Emirates. https://interparestrustai.org/assets/public/dissemination/RA03-Digitization_AI-2023_Feb-Abu_Dhabi_pptx.pdf

Silberztein, M. (2024). The Limitations of Corpus-Based Methods in NLP. In M. Silberztein (Ed.), *Linguistic Resources for Natural Language Processing: On the Necessity of Using Linguistic Methods to Develop NLP Software* (pp. 3–24). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43811-0_1

Singh, S. (2013). Optical Character Recognition Techniques: A survey. *International Journal of Advanced Research in Computer Engineering & Technology*, *2*(6), 2009–2015. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=375e6f94bb9039f3df4fa2a625f11ac59db3629f

Paolanti, M., Pietrini, R., Della Sciucca, L., Balloni, E., Compagnoni, B. L., Cesarini, A., Fois, L., Feliciati, P., & Frontoni, E. (2022). PergaNet: A Deep Learning Framework for Automatic Appearance-Based Analysis of Ancient Parchment Collections. In P. L. Mazzeo, E. Frontoni, S. Sclaroff, & C. Distante (Eds.), *Image Analysis and Processing. ICIAP 2022 Workshops* (Vol. 13374, pp. 290–301). Springer International Publishing. https://doi.org/10.1007/978-3-031-13324-4_25

Pugh, E. (2024). Computer Vision in the Archives. *The Art Bulletin*, *106*(2), 15–18. https://doi.org/10.1080/00043079.2024.2296271

Terras, M. (2022). Chapter 7: Inviting AI into the Archives: The Reception of Handwritten Recognition Technology into Historical Manuscript Transcription. In L. Jaillant (Ed.), *Archives, Access and Artificial Intelligence* (pp. 179–204). Bielefeld University Press. https://doi.org/10.1515/9783839455845-008

Terras, M. M. (2011). The Rise of Digitization. In R. Rikowski (Ed.), *Digitisation Perspectives* (pp. 3–20). SensePublishers. https://doi.org/10.1007/978-94-6091-299-3_1

The National Archives. (n.d.). Born-digital records and metadata [Text]. *The National Archives*. Retrieved February 6, 2025, from https://www.nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/what-are-born-digital-records/

Travis, D. M., Lee, M., Rojas, M., Gunn, A., Nimkar, A., Jansen, G., Diakopoulos, N., & Marciano, R. (2016). Unlocking the Archives of Displacement and Trauma: Revealing Hidden Patterns and Exploring New Modes of Public Access through Innovative Partnerships and Infrastructure. *Archiving Conference*, *13*(1), 135–139. https://doi.org/10.2352/issn.2168-3204.2016.1.0.135

Strien, D. van, Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., & Colavizza, G. (2025). *Assessing the Impact of OCR Quality on Downstream NLP Tasks*. 484–496. https://www.scitepress.org/Link.aspx?doi=10.5220/0009169004840496

Wang, H., Pan, C., Guo, X., Ji, C., & Deng, K. (2021). From object detection to text detection and recognition: A brief evolution history of optical character recognition. *WIREs Computational Statistics*, *13*(5), e1547. https://doi.org/10.1002/wics.1547

Zaagsma, G. (2023). Digital History and the Politics of Digitization. *Digital Scholarship in the Humanities*, *38*(2), 830–851. https://doi.org/10.1093/llc/fqac050