United Nations
Educational, Scientific and
Cultural Organization

Sub-Committee on Education
and Research
Memory of the World

InterPARES
TrustAI

SCEaR Newsletter 2024
Special Issue 2024

Artificial Intelligence and Documentary Heritage

Edited by Luciana Duranti and Corinne Rogers

Editors of this Issue

Dr. Luciana Duranti, Principal Investigator, InterPARES (1998-2026); Professor of archival science, University of British Columbia, Vancouver, Canada; Member of SCEaR, UNESCO Memory of the World Programme.

Luciana.duranti@ubc.ca

Dr. Corinne Rogers, Project Coordinator of InterPARES Trust AI and Adjunct Professor in the School of Information at the University of British Columbia,Vancouver, Canada.

Corinne.rogers@ubc.ca

InterPARES Trust AI is a Cooperating Institution of the Memory of the World Sub-Committee on Education and Research (SCEaR).

Date of delivery of this Issue: 27 May 2024

SCEaR Newsletter
Editors: Lothar Jordan (editor-in-chief), Roslyn Russell, Papa Momar Diop (French texts)

This Newsletter is a medium of the Sub-Committee on Education and Research (SCEaR), IAC, UNESCO Memory of the World Programme.
More information "On SCEaR, Impressum" (p. 55) in
SCEaR Newsletter 2021/2 (December)

Contents

Introduction

by Luciana Duranti and Corinne Rogers

From the early days of machine readable records to the present, each technological advance has brought about excitement for the affordances of new documentary capabilities and documentary forms, as well as concern for how the new technologies would impact the identification, preservation, and trustworthiness of records. Since 1998, InterPARES (International Research on the Preservation of Authentic Records in Electronic Systems), a multinational and multidisciplinary collaborative research project funded by the Social Sciences and Humanities Research Council of Canada and the University of British Columbia (UBC), has challenged these concerns and shown, through five phases of research, that archival principles and concepts can support the trustworthiness of records regardless of the technology used to create, manage, and preserve them. Today Artificial Intelligence (AI) is changing the way we live and work, sparking either great enthusiasm for its capabilities or great fear for its unintended consequences.[1] While AI poses significant challenges to documentary heritage, these challenges can be met with the same grounding in foundational archival knowledge when approached through collaborative, interdisciplinary effort.

This Special Issue of the *SCEaR Newsletter* focuses on the uses and impacts of AI in the field of documentary and cultural heritage. The articles that follow offer a glimpse of some of the research underway through InterPARES Trust AI (I Trust AI), the fifth phase of InterPARES.[2]

I Trust AI includes almost 100 partner organizations in North, Central, and South America, Europe, Africa and Asia, representing academia, industry, archives, government agencies, and international organizations. Researchers from the archives and records fields, cybersecurity, data science, information science, computer science, computational linguistics, economics, law, electrical engineering and more are engaged in more than 40 studies investigating the current and potential uses of AI in records and archival work, focusing attention on the challenges to trustworthiness of documentary heritage arising from rapid and ill-considered adoption of AI.

These articles present a broad overview of the type of research underway in I Trust AI. Space constraints dictated, however, that only a few studies be presented, reflecting the breadth of research and the internationality and multidisciplinarity of the researchers and authors. Of greatest importance through all the studies has been and continues to be the successful establishment of working relationships and mutual understanding between archives and records specialists and AI scientists. A terminology database supports such understanding by defining terms and concepts used by each discipline. In addition, the I

[1] For current public opinion about the capabilities and risks of AI see *Artificial Intelligence Index* (2023), Stanford Institute for Human-Computer Interaction, https://aiindex.stanford.edu/report/.

[2] I Trust AI, the fifth phase of InterPARES, was introduced in the *SCEaR Newsletter* 2022/1 (June) in Luciana Duranti's article "Artificial Intelligence for Documentary Heritage", p. 11. See also https://interparestrustai.org.

Trust Team is working with ISO TC 46/SC 11 on terminology, and with the Canadian General Standards Board (CGSB) on the third edition of the *Electronic Records as Documentary Evidence* standard, by bringing AI into it.

In the first article Dr. Muhammad Abdul-Mageed, Canada Research Chair in Natural Language Processing and Deep Learning and, with Luciana Duranti, co-director of I Trust AI, sets the stage, offering an overview of the AI research underway through the work carried out at the NLP (Natural Language Processing) lab at UBC, and presenting some of the vast capabilities of AI to assist in serving diverse and under-represented communities. The application of these technologies to archives is discussed, and the ethical implications identified. This is followed by an article by Nagoudi et al. that focuses on the tendency of large language models (LLM) to generate "hallucinations" – content that is fabricated and false – and presents a developing solution called Retrieval Augmented Generation (RAG) to enhance the reliability of LLMs and decrease the occurrence of hallucinations.

The UNESCO audio archives (Sullivan and Sengsavang) offers a case study of some of the archival challenges – language identification in audio recordings and metadata enrichment to enhance description and retrieval, that may be alleviated using AI technologies. Another significant challenge – the identification of personally identifiable information and protection of personal privacy in archival holdings – is addressed by Sullivan et al. A different approach to the protection of privacy is discussed by Lemieux, who introduces privacy enhancing technologies. Allegrezza et al. present their research into classification, aggregation or reaggregation of records using AI tools.

The bridge between the 14th and the 21st century is being built using deep neural networks. Frontoni introduces his AI-driven system, PergaNet, to analyse medieval parchments, a very large quantity of which languishes in European archives and now may be identified.

The trustworthiness of AI outputs, as well as the transparency of and accountability for AI use, rely on the generation and preservation of paradata. This is discussed in the articles by Franks and Cameron.

How AI is incorporated into the archival workflow is the subject of the next two articles. Sengsavang and Trbušić present a model for AI-assisted digitization, developed for the UNESCO archives. Stančić and Trbušić discuss archival challenges and an AI workflow more generally. This is followed by Bushey's article considering the challenges and implications of generative AI on images archives.

Guerrero et al. present a study that investigates the comprehensiveness of Sweden's, Finland's, and South Africa's existing and emerging regulatory frameworks for AI and identify gaps in them regarding ongoing trustworthiness of public records. Finally, Rockembach argues for the need for AI literacy among records and archives professionals.

The researchers in the NLP lab at UBC are also developing tutorials for archival and broader use in natural language processing, part-of-speech tagging, named entity recognition, text classification, machine translation, and automatic speech recognition. More tutorials are planned in image processing and practical machine learning. All are or will be freely available at https://github.com/UBC-NLP/I Trust AI-tutorials.

Digital Twins, mentioned in Cameron's article in relation to the need for paradata, are the subject of research led by I Trust AI partner, Carleton University (Ottawa, Canada). The Carleton team is investigating the following questions: Can a digital twin be preserved, and what is required at the point of creation to ensure that it can be? Can the AI, automation and real time data involved in this complex social and technological type of system be preserved? What might be the role of AI be in terms of creating an archival package to ingest a digital twin?

As mentioned by Abdul-Mageed, generative AI leverages artificial intelligence to produce synthetic multimedia documents, such as text, images, audio and video. While multimedia generally encompasses video and audio content, mulsemedia (multiple sensorial media) also includes haptic, gustatory, olfactory, and media including more than two senses. Massive production of AI-generated multimedia and mulsemedia data is expected in the upcoming years. Veracity and truth assessment tools, such as AI-generated detectors, will become increasingly critical for archivists and users of archives in the current and evolving context of digital data. The field of records and archives currently lacks the tools to assess vast volumes of multimedia data. A new I Trust AI study is beginning to explore various aspects of data veracity, truth discovery of multimedia data, and the challenges emerging due to the novel AI generative techniques, in order to identify promising research directions and approaches. It is expected that the results of this study will have implications for software vendors, archives and all those organizations that must work with large volumes of multimedia data the veracity of which is a concern.

Following is a list of some of the more commonly used acronyms pertaining to Artificial Intelligence found in this newsletter:

AI	Artificial Intelligence
DL	Deep Learning
HTR	Handwritten Text Recognition
HWR	Handwriting Recognition
LLM	Large Language Model
MLLM	Multimodal Large Language Model
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OCR	Optical Character Recognition
RAG	Retrieval Augmented Generation
VQA	Visual Question Answering
WER	Word Error Rate

We hope that the readers will enjoy reading this Special Issue of the *SCEaR Newsletter*, and are encouraged and inspired by the work underway to understand the impacts and uses of AI in the documentary heritage fields.

And finally we thank Lothar Jordan, Chair of the MoW SCEaR, for the good cooperation and his editorial support.

Dr. Luciana Duranti, Principal Investigator of I Trust AI, is Professor of Archival Science at the University of British Columbia, Canada, Affiliate Full Professor at the University of Washington at Seattle, USA, and Chair of the Canadian Government Standards Board Committee for Electronic Records as Documentary Evidence. She has been a member of the UNESCO International Advisory Committee (IAC) of the Memory of the World Programme and the organizer of the 2012 Vancouver Conference "The Memory of the World in the Digital Age: Digitization and Preservation," which issued the "MoW Vancouver Declaration." She is a member of the MoW SCEaR.

Dr. Corinne Rogers is the Project Coordinator for InterPARES Trust AI (UBC, 2021-2026), and previously InterPARES Trust (UBC, 2012-2019). She is an adjunct professor in the Information School at the University of British Columbia on the subjects of diplomatics, digital records forensics, and digital preservation. She is Co-Convenor of the Working Group on Electronic Records as Documentary Evidence, Canadian General Standards Board.

AI in the I Trust AI Partnership

by Muhammad Abdul-Mageed

Introduction

In the *I Trust AI* partnership our endeavours span a broad spectrum. On the one hand, *training high-quality students, postdoctoral fellows, fellow faculty members, and professionals* in the field is one of our top priorities. A notable challenge in this respect that we have identified is bridging the communication gap between diverse groups. After all, aligning the discourse between a computer science student and a city archivist, for example, requires time and efforts. To this end, we have diligently organized plenaries, symposia, and webinars. We have also delivered invited talks, guest lectures, and hands-on workshops. These efforts aim to provide the necessary training to forge a critical mass capable of engaging in the interdisciplinary work essential to our mission. On the other hand, there is an ongoing need to *develop cutting-edge technologies* that not only meet the current requirements of archivists but also anticipate future demands. A significant hurdle in this pursuit has been the time required to gather datasets from actual end users for technology development. To overcome this challenge, we have pursued a dual-track approach: (1) encouraging professionals to identify and develop usable datasets for technology advancement, and (2) utilizing available public datasets to enable the rapid development of technologies that can later be tailored to the specific needs of our project partners.

Projects within *I Trust AI* are quite diverse. Some projects aim to comprehend how technology is presently perceived, needed, or utilized in specific professional contexts. Others adopt a more practical stance, often involving collaborations with AI professionals to repurpose existing technologies for identified needs. A third category is dedicated to identifying challenges that can be addressed with new technologies, employing innovative methods designed to demonstrate potential within the broader partnership. This last category allows for the insights and software developed to be applied to pertinent needs later.

As the co-director of the partnership, my responsibilities span these different projects, involving different levels of engagement. My role ranges from providing answers and resources, offering high-level guidance, to actively participating in given projects. In this article, I will offer an overview of selected projects conducted in my own research group. These projects, primarily led by my graduate students and postdoctoral fellows, vary in their applicability to both national and international archives in the sense that some of them are developed with actual archives datasets to serve one or more partners while others have broader applicability and exploit general public data. Many of these projects are based on language models, and the most recent among them lean towards *Generative AI* research.

In the following sections, I will first introduce Generative AI as an area of research and development, followed by a cluster of works focused on serving diverse communities, Arabic and African languages, and massively multilingual contexts. I will then highlight a number of initial efforts aiming at management of private archives, multimodal content, archives in need of optical character recognition, and speech processing. I conclude with a note on ethical considerations.

Generative AI

The term *Generative AI*, as it is currently understood, primarily describes machine learning models and methodologies adept at generating content in response to user prompts. This encompasses a broad range of outputs, including audio, images, text, and video. The high fidelity of this AI-generated content has recently captivated users, with machines capable of producing content that is often indistinguishable from that created by humans. At the core of Generative AI are foundation models, often referred to as large language models (LLMs) in the context of text generation, which are pretrained on vast datasets. For instance, LLMs have been pretrained on extensive collections of web texts, books, and social media posts. Some proprietary models, such as ChatGPT, have additionally been pretrained on licensed datasets. It is commonly observed that larger models, which are pretrained on more extensive datasets, outperform their smaller counterparts limited to narrower data scopes. The significance of datasets in the development of these models cannot be overstated. Other critical factors in the development of these models include technical expertise and a clear vision to customize these models for specific use cases and to meet the requirements of particular user communities. Successful work in Generative AI is hardly the result of direct development of models. In other words, there are usually needs to build codebases for collecting datasets in particular languages, which in turn require access to mature language processing and identification technologies. I will now introduce a number of studies focused at improving access to archives in multilingual content that showcase these types of nuanced needs for a wide range of languages.

Serving Diverse Communities

Expanding technological reach to archives in various languages presents a significant challenge. The limitation of existing AI technologies in terms of language coverage is stark; a vast majority of the world's 7000+ languages are not supported. This gap is particularly concerning as technology becomes increasingly integral to archiving, potentially exacerbating inequalities and biases. The release of ChatGPT, accompanied by widespread excitement about its capabilities, prompted us to question the breadth of languages the model could recognize and serve effectively. To explore ChatGPT's language identification capabilities, we conducted a study titled "Fumbling in Babel: An Investigation into ChatGPT's Language Identification Ability" (Chen et al., 2024). Our research introduced *Babel-670*, a benchmark encompassing 670 languages across 23 language families and five continents, ranging from highly to minimally resourced languages. Our examination of ChatGPT's performance (across versions GPT-3.5 and eGPT-4) focused on its ability to: (i) identify language names and codes; (ii) operate under zero- and few-shot conditions; and (iii) function with or without a provided label set. The findings revealed that, while the model could accurately identify nearly a hundred languages (achieving an F_1 score above 90%)[3], it showed negligible knowledge for another 382 languages, where it nearly failed to achieve any F_1 score. Geographically, African languages received the least support from ChatGPT (see *Figure 1*). Compared to specialized, fine-tuned language identification tools, ChatGPT was found to be lacking. These results highlight a critical issue: ChatGPT, and likely most existing large language models (LLMs), fall short of serving the wide and diverse linguistic needs of global communities in their native languages.

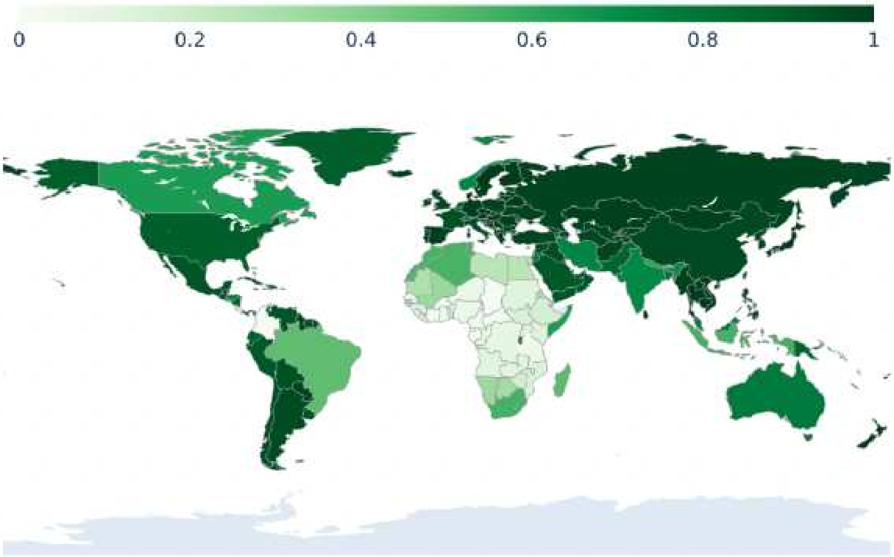


Figure 1. A choropleth map where the intensity indicates the averaged F_1 score of languages spoken in each region. It can be seen that the support of languages has geographical discrepancy, e.g. with African languages being strikingly less supported. Photo: Chen et al. (2024).

To address the challenges in identifying African languages, we developed AfroLID, a neural language identification (LID) toolkit designed for 517 African languages and dialects, as detailed in another study (Adebara et al., 2022; see *Figure 2*). AfroLID leverages a manually curated, multi-domain web dataset drawn from 14 language families and incorporating five orthographic systems. In blind testing, AfroLID achieved a remarkable 95.89 F_1 score. Our comparisons with five existing LID tools, each covering a limited subset of African languages, demonstrated AfroLID's superior performance across the majority of languages assessed. Additionally, we validated AfroLID's practicality in real-world applications by deploying it to analyze content from the significantly underserved Twitter domain. Further exploration through controlled case studies and a linguistically driven error analysis illuminated both the strengths and

[3] F_1 score is a measure used to test the accuracy of a model. It considers both the precision of the model (how many identified items are actually correct) and its recall (how many of the correct items it can identify). A higher F_1 score means the model is more accurate.

limitations of AfroLID. Subsequently, we packaged AfroLID as a standalone Python package, making it publicly available and creating a web demo for interactive experimentation with the underlying model. AfroLID serves as a paradigmatic example of how to develop LID technologies for a broader range of languages. Only by expanding the coverage of natural language processing tools can we cater to diverse use cases and better serve archival communities worldwide.

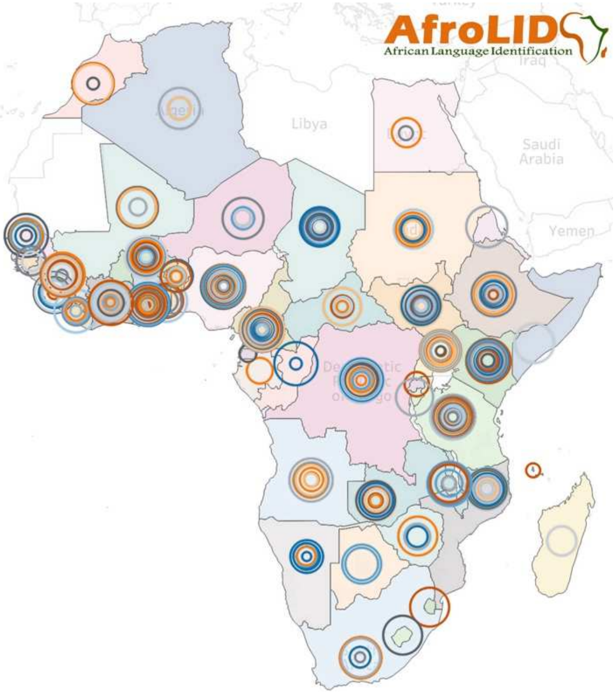


Figure 2. All 50 African countries in AfroLID data, with 517 languages/language varieties in colored circles overlayed within respective countries. Photo: Adebara et al. (2022).

Natural Language Processing and Archives

Language identification, while crucial, is merely the first step in real-world work needed for archives. To address this, we have initiated a series of targeted studies aimed at broadening the technological capabilities for a diverse array of languages across numerous tasks in both *natural language understanding (NLU)* and *natural language generation (NLG)*. NLU focuses on enabling machines to comprehend and interpret human (natural) language, tackling tasks such as understanding the meanings of words, phrases, sentences, and larger text segments. This facilitates functions like named entity recognition (identifying names of persons, places, organizations within a text), sentiment analysis (assessing the emotional tone of a text, whether neutral, positive, negative, or mixed), and part of speech tagging (determining the grammatical category of words or sub-words, such as nouns, verbs, adjectives, and adverbs). NLG, conversely, focuses on enabling machines to produce coherent language sequences for tasks like summarization (creating concise summaries from longer texts), machine translation (translating non-English languages into accurate, faithful, and fluent English), and question answering (providing satisfying and precise responses to queries).

NLP capabilities are particularly valuable for *archival applications*. For example, named entity recognition (NER) systems can enable classifying records and arrange and describe archival fonds based on the names of people, places, and organizations, thus improving description of these materials and facilitating access to them. Some technologies can be used to summarize records or archival fonds content, translate it from one language to another, and/or enable users to ask questions about content and receive tailored answers. These technologies can thus significantly enhance engagement of users with the archival content and reduce human effort needed to locate particular types of information. Our NLP efforts are directed towards developing advanced language models that offer comprehensive solutions in these domains, often spanning multiple languages. I introduce some of these next.

Arabic NLP

In this research, we focus on developing models and applications specifically for the Arabic language. It is crucial to understand that Arabic encompasses a wide array of languages, language varieties, and dialects, serving as the mother tongue for over 450 million people predominantly in Africa and Asia. Our models are designed to cater to the modern variant, Modern Standard Arabic (MSA), the classical form known as Classical Arabic (CA), as well as various country-level dialects.

In Abdul-Mageed et al. (2021) we present two deep bidirectional transformer-based models for Arabic NLU: ARBERT and MARBERT. Alongside these models, we introduce ARLUE, a comprehensive new benchmark designed for evaluating multi-dialectal Arabic language understanding. ARLUE incorporates 42 datasets targeting six distinct task clusters, enabling us to conduct a series of standardized tests under diverse conditions. Upon fine-tuning with ARLUE, our models achieved state-of-the-art (SoTA) results in the majority of tasks, specifically in 37 out of 48 classification tasks across the 42 datasets. Notably, our leading model attained the highest ARLUE score (77.40) among all task clusters, surpassing all competing models, including the significantly larger XLM-R-Large. We have made our models publicly accessible at https://github.com/UBC-NLP/marbert, facilitating further research and application development in Arabic NLU.

Following our work on NLU, we have developed a suite of encoder-decoder models for Arabic NLG, named AraT5. These new models leverage the unified Transformer framework (T5), transforming all language-related problems into a text-to-text format. This approach embodies a straightforward yet powerful method for transfer learning. To assess the model's effectiveness, we introduce a pioneering benchmark specifically designed for Arabic language generation (ARGEN), encompassing seven critical tasks (see *Figure 3*). For a comprehensive evaluation, we pretrain three robust variants of our Arabic T5-style models and benchmark them against ARGEN. Despite being pretrained with approximately 49% less data, our new models outperform the multilingual mT5 from Google across all ARGEN tasks (achieving superior results in 52 out of 59 test sets) and establish several new SoTAs. Again, we make our models publicly accessible on GitHub: https://github.com/UBC-NLP/araT5.

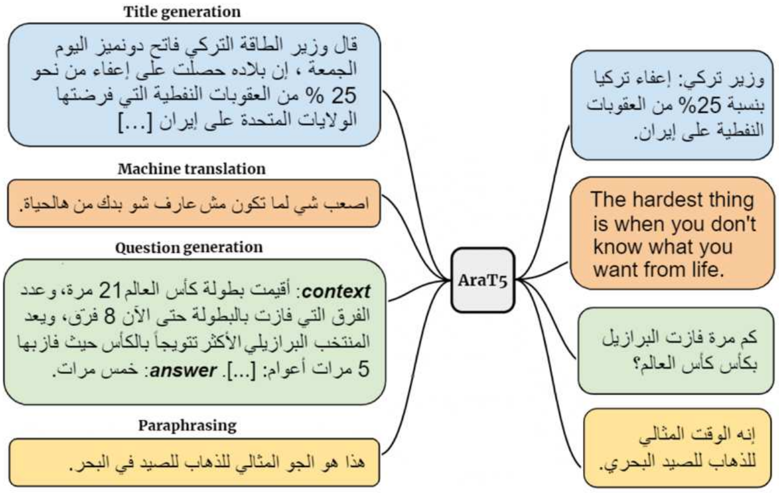

Title generation
قال وزير الطاقة التركي فاتح دونميز اليوم الجمعة ، إن بلاده حصلت على إعفاء من نحو 25 % من العقوبات النفطية التي فرضتها الولايات المتحدة على إيران [...]
Machine translation
اصعب شي لما تكون مش عارف شو بدك من هالحياة.
Question generation
context: أقيمت بطولة كأس العالم 21 مرة، وعدد الفرق التي فازت بالبطولة حتى الآن 8 فرق، ويعد المنتخب البرازيلي الأكثر تتويجاً بالكأس حيث فازبها 5 مرات أعوام: [...]. answer: خمس مرات.
Paraphrasing
هذا هو الجو المثالي للذهاب للصيد في البحر.
AraT5
وزير تركي: إعفاء تركيا بنسبة 25% من العقوبات النفطية على إيران.
The hardest thing is when you don't know what you want from life.
كم مرة فازت البرازيل بكأس كأس العالم؟
إنه الوقت المثالي للذهاب للصيد البحري.

Figure 3. AraT5 encoder-decoder model and prompt samples from four investigated tasks, namely: title generation, machine translation, question generation, and paraphrasing. Photo: Nagoudi et al. (2022).

Our research on Arabic, akin to our investigations into African languages, offers numerous potential applications for other linguistically under-resourced languages. For instance, researchers aiming to develop technological solutions for these languages might adopt our strategies for dataset collection and model development. Similarly, our approaches to model evaluation can be tailored for use with these languages. Furthermore, the technologies we have developed for Arabic could be effectively integrated into projects with multilingual goals. Given its diverse array of dialects and broad geographic distribution, Arabic provides an exemplary case for extending these methodologies to additional languages. Consequently, our work in Arabic NLP can be intimately relevant for contexts requiring technological support for non-Arabic archives. We are also currently extending work we have carried out on Arabic to several other languages, as part of the *I Trust AI* partnership.

African NLP

Multilingual pretrained language models have been instrumental in capturing valuable, generalizable linguistic knowledge during their pretraining phase, significantly pushing forward the SoTA benchmarks through task-specific finetuning. However, a notable gap exists in their coverage, with only approximately 31 out of 2,000 African languages represented in these models to date. To address this shortfall and allow for widening work on African archives, we introduce SERENGETI (Adebara et al., 2023), a comprehensive multilingual language model designed to encompass 517 African languages and dialects. Our models were assessed across eight NLU tasks utilizing 20 distinct datasets, in comparison with four existing multilingual pretrained models, each of which covers between 4 to 23 African languages. In this rigorous evaluation, SERENGETI demonstrated superior performance on 11 datasets spanning all eight tasks, achieving an impressive average F_1 score of 82.27. Additionally, we conducted an error analysis to explore the impact of language genealogy and linguistic similarity on model performance, particularly in zero-shot settings (i.e. settings where the model can learn to recognize patterns which it has not been explicitly trained on*)*. In the spirit of advancing research and fostering further advancements, we have made our SERENGETI models publicly available to the research community at https://github.com/UBC-NLP/serengeti.

We also built models for African NLG. In Adebara et al. (2024), we develop Cheetah, a massively multilingual NLG language model for African languages. Cheetah supports 517 African languages and language varieties, allowing us to address the scarcity of NLG resources and provide a solution to foster linguistic diversity. Our evaluation of Cheetah spans seven downstream generation tasks (see *Figure 4*), where it significantly surpasses existing models in five tasks. This underscores Cheetah's exceptional ability to produce coherent and contextually relevant text across a diverse spectrum of African languages. Further insights were gained through an in-depth human evaluation, assessing Cheetah's linguistic proficiency. Cheetah's development marks a significant stride towards enhancing linguistic diversity, offering a scalable method to adapt pretrained models for specific languages. This facilitates the creation of viable NLG applications for African communities, contributing significantly to NLP advancements in resource-scarce environments. Such efforts are crucial for increasing accessibility and ensuring the inclusion of African languages in the burgeoning digital domain.

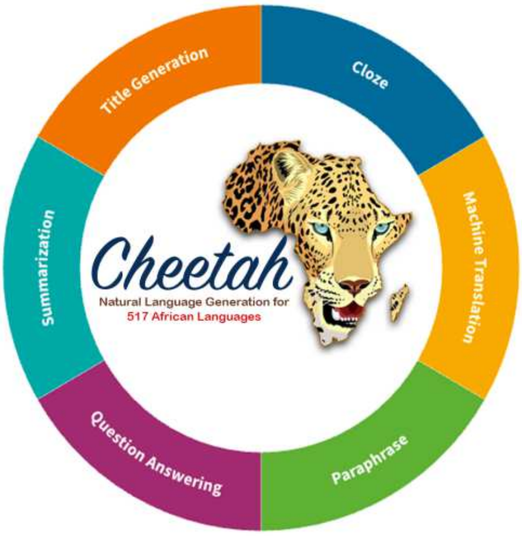


Figure 4. Cheetah is trained on 517 African languages and language varieties across 14 language families. The languages are domiciled in 50 out of the 54 African countries and are written in six different scripts. Photo: Adebara et al. (2024).

Massively Multilingual NLP

In Zhang et al. (2023) we examine the performance of instruction-tuned LLMs such as ChatGPT on cross-lingual sociopragmatic meaning (SM). SM refers to meaning embedded within social and interactive contexts. To appreciate SM, consider how the meaning of an utterance in social interaction (e.g., on social media) can be highly subtle and how it incorporates both the social variation related to language users (from a sociolinguistics perspective) and their communicative intentions (from a pragmatics perspective). Although SM is quite established within linguistics, NLP systems still struggle with this type of meaning and there is a gap in studying it. This deficiency arises partly from SM not being adequately represented in any of the existing benchmarks. To address this gap, we present SPARROW, an extensive multilingual benchmark specifically designed for SM understanding. SPARROW comprises 169 datasets covering 13 task types across six primary categories (e.g., anti-social language detection, emotion recognition). SPARROW datasets encompass 64 different languages originating from 12 language families representing 16 writing scripts. We evaluate the performance of various multilingual pretrained language models (e.g., mT5) and instruction-tuned LLMs (e.g. BLOOMZ, ChatGPT) on SPARROW through fine-tuning, zero-shot, and/or few-shot learning. Our comprehensive analysis reveals that existing open-source instruction tuned LLMs still struggle to understand SM across various languages, performing close to a random baseline in some cases. We also find that although ChatGPT outperforms many LLMs, it still falls behind task-specific finetuned models with a gap of 12.19 SPARROW score. We also offer highly effective smaller models for the SM in our benchmark. Our benchmark is available at: https://github.com/UBC-NLP/SPARROW. There is often a need to *understand the sociopragmatics of archival texts*, for example, identifying emotions accompanying certain events or towards particular topics or individuals. This work paves the way to applications involving sociopragmatic models in these contexts.

Multimodal and OCR Models

Various archival institutions have identified a crucial need for models that can perform multimodal understanding, where a model processes an image to extract specific types of information or answer questions about objects within that image. Classic examples include optical character recognition (OCR) and handwriting recognition (HWR). In response to this need, our ongoing research introduces *Qalam*, which means "pen" in Arabic. Qalam is a pioneering foundation model tailored to enhance Arabic OCR and HWR capabilities (see *Figure 5* and *Figure 6*). It utilizes a cutting-edge SwinV2 encoder and RoBERTa decoder architecture, achieving outstanding performance across various datasets, highlighted by remarkably low Word Error Rate (WER) scores in HWR (0.80%) and OCR (1.18%) tasks. Qalam distinguishes itself through advanced pretraining strategies and data augmentation techniques, specifically addressing the challenges associated with Arabic script recognition such as diacritic handling and high-resolution input scalability. The model notably surpasses current SoTA solutions, demonstrating the impactful potential of transformer-based models in improving OCR and HWR systems.

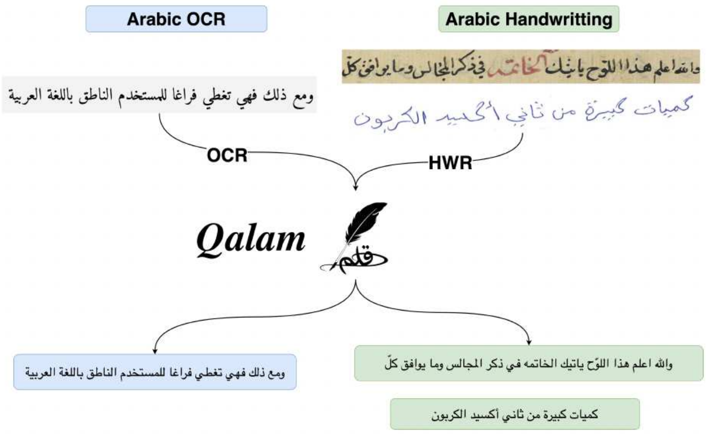
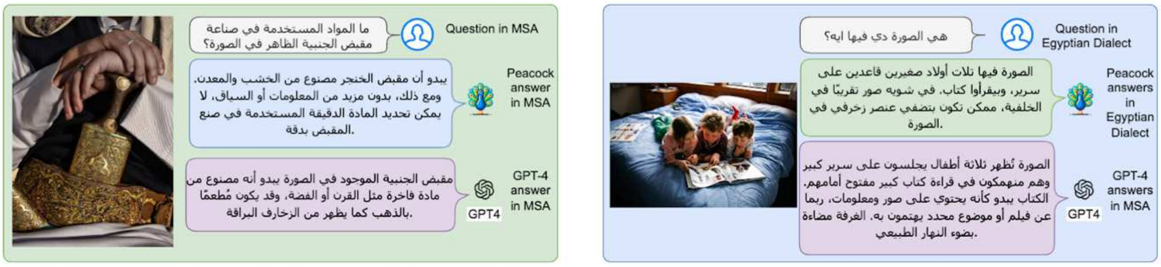

Arabic OCR
Arabic Handwritting
ومع ذلك فهي تغطي فراغا للمستخدم الناطق باللغة العربية
والله اعلم هذا اللوح ياتيك الخاتمه في ذكر المجالس وما يوافق كل
كميات كبيرة من ثاني أكسيد الكربون
OCR
HWR
Qalam
قلم
ومع ذلك فهي تغطي فراغا للمستخدم الناطق باللغة العربية
والله اعلم هذا اللوّح ياتيك الخاتمه في ذكر المجالس وما يوافق كلّ
كميات كبيرة من ثاني أكسيد الكربون

Figure 5. An illustrative overview of *Qalam* on arabic OCR and HWR across diverse text types. Photo: Bhatia et al. (in preparation, 2024).

Figure 6. Historical manuscript data samples from the *Qalam* study. Photo: Bhatia et al. (in preparation, 2024).

In Alwajih et al. (2024) we extend our multimodal efforts by introducing *Peacock*, a comprehensive family of Arabic multimodal LLMs (MLLMs) developed to address the scarcity of high-quality multimodal resources for languages other than English, particularly for Arabic. Despite Arabic's widespread use, the success of MLLMs has been largely limited to English due to the lack of quality multimodal resources. Peacock aims to bridge this gap by providing a robust collection of models with advanced vision and language capabilities. Architecture of Peacock models combines an image encoder with an Arabic text decoder, leveraging both a fully connected layer and a Q-former transformer for aligning vision and language components. The models undergo a two-stage training process, first on high-quality pre-training data translated into Arabic and then on instruction finetuning datasets, to enhance their reasoning and conversational abilities. We also introduce *Henna*, a benchmark specifically designed for assessing MLLMs on aspects related to Arabic culture, setting a foundation for culturally aware Arabic MLLMs. Peacock models excel in visual question answering (VQA) and visual reasoning tasks, showing a significant improvement when trained on filtered high-quality data (see *Figure 7*).



ما المواد المستخدمة في صناعة مقبض الجنبية الظاهر في الصورة؟

يبدو أن مقبض الخنجر مصنوع من الخشب والمعدن. ومع ذلك، بدون مزيد من المعلومات أو السياق، لا يمكن تحديد المادة الدقيقة المستخدمة في صنع المقبض بدقة.

مقبض الجنبية الموجود في الصورة يبدو أنه مصنوع من مادة فاخرة مثل القرن أو الفضة، وقد يكون مُطعمًا بالذهب كما يظهر من الزخارف البراقة.

هي الصورة دي فيها ايه؟

الصورة فيها تلات أولاد صغيرين قاعدين على سرير، وبيقرأوا كتاب. في شويه صور تقريبًا في الخلفية، ممكن تكون بتضفي عنصر زخرفي في الصورة.

الصورة تُظهر ثلاثة أطفال يجلسون على سرير كبير وهم منهمكون في قراءة كتاب كبير مفتوح أمامهم. الكتاب يبدو كأنه يحتوي على صور ومعلومات، ربما عن فيلم أو موضوع محدد يهتمون به. الغرفة مضاءة بضوء النهار الطبيعي.

Figure 7. Examples of responses from Peacock and GPT-4V regarding an image related to Yemeni culture (left). Both Peacock and GPT-4V accurately respond to a question in the Egyptian dialect. While GPT-4V provides a slightly more detailed answer, it does so in MSA. In contrast, Peacock's response is in the same Egyptian dialect as the question (right). Photo: Alwajih et al. (2024).

Speech Processing

Some of our other ongoing work involves developing speech processing models that can also be adapted to applications in archives, building on our own previous work on automatic speech processing (Talafha et al., 2023) and speech dialect identification (Sullivan et al., 2023). One application domain is the UNESCO archives where we work with our partners to facilitate access and organization of their audio archives. You will find a more detailed account of this work in another article in this issue of the *SCEaR Newsletter* (Sullivan and Sengsavang, pp. 27-33).

Private and Personal Information in Archives

Our work also encompasses projects with a narrower focus, addressing specific archival challenges. Work on management of *private and personal information (PII)* is an example. PII encompasses any data that could potentially identify, contact, or locate an individual, either on its own or when combined with other personal or identifying data that is linked or linkable to a specific individual. This category includes direct identifiers, such as social security numbers, email addresses, and phone numbers, as well as indirect identifiers. The latter, when aggregated, could identify an individual through a combination of attributes like gender, race, birthdate, and geographical indicators. The careful handling of PII underscores the broad archival need to balance the release of information with the imperative to protect individual privacy and security. A particular project we are conducting as a forerunner of PII management is related to text rewriting in order to remove toxicity, a task known as "detoxification". Previous detoxification research has been fragmented, focusing on limited platforms without considering real-world diversity or the challenge of non-detoxifiability, where detoxification changes the text's meaning.

In Khondaker et al. (2024) we introduce GreenLLaMA, a holistic end-to-end detoxification framework designed to overcome these issues (see *Figure 8*). It includes a novel cross-platform pseudo-parallel corpus created with advanced data processing and generation techniques using ChatGPT. Our models, trained on this corpus, surpass SoTA models in detoxifying content across various platforms while maintaining the original meaning. GreenLLaMA enhances transparency with explanations for its detoxification decisions and incorporates a paraphrase detector to address non-detoxifiable content, signaling when meaning might be altered. This approach not only showcases GreenLLaMA's superior detoxification capabilities but also its resilience against adversarial attacks, offering a practical solution for real-world application.

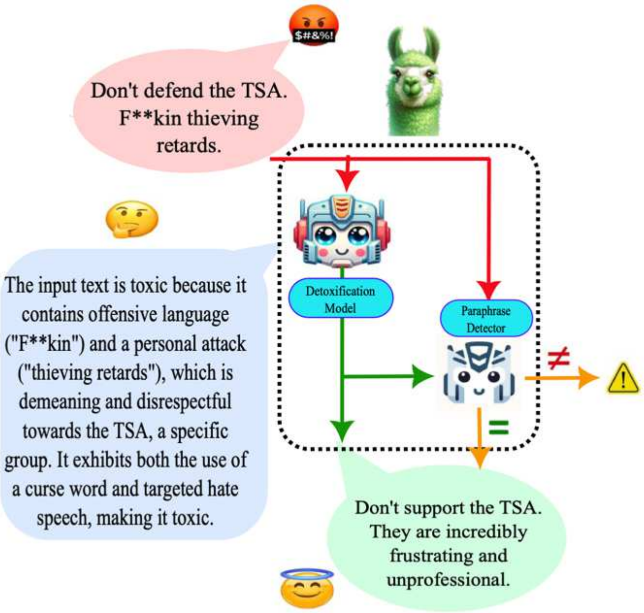

$#&%!
Don't defend the TSA. F**kin thieving retards.
The input text is toxic because it contains offensive language ("F**kin") and a personal attack ("thieving retards"), which is demeaning and disrespectful towards the TSA, a specific group. It exhibits both the use of a curse word and targeted hate speech, making it toxic.
Detoxification Model
Paraphrase Detector
≠
=
Don't support the TSA. They are incredibly frustrating and unprofessional.

Figure 8. Workflow of GreenLLaMA framework. The framework will take a toxic input. The detoxification model will generate the explanation of why the input is toxic, as well as provide a non-toxic version. The paraphrase detector will analyze the semantic similarity of the toxic and non-toxic pair and generate a warning if the two are not similar in meaning. Photo: Khondaker et al. (2024).

Ethics in *I Trust AI*

In parallel to our efforts to develop new technologies for addressing challenges in archives, we depend on a broad network of partners and collaborators to foster discussions about ethical considerations and policy implications as these technologies are adopted more widely. While several ongoing studies are examining the ethics of using AI in archives, these topics extend beyond the scope of the current article. However, it is important to note that each study presented here includes a section on limitations of the technology presented and related ethical considerations. Maintaining a focus on these ethical aspects is crucial throughout our work.

Conclusion

Overall, our methodology in developing AI technologies for archives under the *I Trust AI* partnership is marked by agility, interdisciplinarity, and inclusivity. We acknowledge the vast diversity within our partnership and understand the significance of employing comprehensive approaches. Consequently, while this article highlights several research-oriented projects, numerous other studies within the partnership are exploring various AI methodologies or concentrating on the application of existing technologies. These studies play a crucial role within our partnership by providing valuable contexts for training both students and professionals, as well as addressing real-world needs. It is through embracing this inclusive and diverse strategy that *I Trust AI* continues to flourish.

References

Abdul-Mageed, M., & Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7088-7105). https://aclanthology.org/2021.acl-long.551.pdf

Adebara, I., Elmadany, A., & Abdul-Mageed, M. (2024). Cheetah: Natural Language Generation for 517 African Languages. *arXiv preprint arXiv:2401.01053*. https://arxiv.org/pdf/2401.01053.pdf

Adebara, I., Elmadany, A., Abdul-Mageed, M., & Inciarte, A. A. (2023, July). SERENGETI: Massively Multilingual Language Models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 1498-1537). https://aclanthology.org/2023.findings-acl.97.pdf.

Adebara, I., Elmadany, A., Abdul-Mageed, M., & Inciarte, A. (2022, December). AfroLID: A Neural Language Identification Tool for African Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 1958-1981). https://aclanthology.org/2022.emnlp-main.128.pdf

Alwajih, F., Nagoudi, E. M. B., Bhatia, G., Mohamed, A., & Abdul-Mageed, M. (2024). Peacock: A Family of Arabic Multimodal Large Language Models and Benchmarks. *arXiv preprint arXiv:2403.01031*. https://arxiv.org/pdf/2403.01031.pdf

Chen, W. R., Adebara, I., Doan, K. D., Liao, Q., & Abdul-Mageed, M. (2023). Fumbling in Babel: An Investigation into ChatGPT's Language Identification Ability. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. *arXiv preprint arXiv:2311.09696*. https://arxiv.org/pdf/2311.09696.pdf

Khondaker, M. T. I., Abdul-Mageed, M., & Lakshmanan, L. V. (2024). GreenLLaMA: A Framework for Detoxification with Explanations. *arXiv preprint arXiv:2402.15951*. https://arxiv.org/pdf/2402.15951.pdf

Nagoudi, E. M. B., Elmadany, A., & Abdul-Mageed, M. (2022). AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers) (pp. 628-647). https://aclanthology.org/2022.acl-long.47.pdf

Sullivan, P., Elmadany, A., & Abdul-Mageed, M. On the robustness of Arabic speech dialect identification. In *Proceedings of the Annual Conf. of the Intl. Speech Communication Association, INTERSPEECH 2023*, pp. 5326-5330, Aug 2023. doi:10.21437/interspeech.2023-1044. https://www.isca-archive.org/interspeech_2023/sullivan23_interspeech.pdf

Talafha, B., Waheed, A., & Abdul-Mageed, M. (2023). N-shot benchmarking of whisper on diverse Arabic speech recognition. In *Proceedings of the Annual Conf. of the Intl. Speech Communication Association, INTERSPEECH 2023*, pp. 5092-5096, Aug 2023. doi:10.21437/interspeech.2023-1044. https://www.isca-archive.org/interspeech_2023/talafha23_interspeech.pdf

Zhang, C., Doan, K., Liao, Q., & Abdul-Mageed, M. (2023). The Skipped Beat: A Study of Sociopragmatic Understanding in LLMs for 64 Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2630-2662). https://aclanthology.org/2023.emnlp-main.160.pdf

Dr. Muhammad Abdul-Mageed is a Canada Research Chair in Natural Language Processing and Machine Learning, and Associate Professor with appointments in the School of Information, and the Departments of Linguistics and Computer Science at The University of British Columbia. He is director of the UBC Deep Learning & NLP Group, co-director of the SSHRC-funded I Trust Artificial Intelligence, and co-lead of the Ensuring Full Literacy Partnership. He is a founding member of the UBC Center for Artificial Intelligence Decision making and Action and a member of the Institute for Computing, Information, and Cognitive Systems.

Improving Archives-Focused LLMs with Retrieval Augmented Generation

by ElMoatez Billah Nagoudi, Alcides Alcoba Inciarte, Abdul-Mageed Muhammad

Introduction

The growth of Large Language Models (LLMs) such as ChatGPT has revolutionized the field of Natural Language Processing (NLP), offering unprecedented capabilities in text generation and understanding. However, a significant challenge faced by these models is their tendency to produce "hallucinated" content – information that is inaccurate or entirely fabricated. This phenomenon not only impacts the reliability of LLM outputs but also limits their application in critical domains requiring high accuracy, such as medical advice or scientific research.

With the development of methods to enhance the reliability of LLMs and minimize the occurrence of hallucinations, a promising solution has emerged in the form of retrieval augmented generation (RAG). This technique enhances LLMs by equipping them with a way to pull in relevant and timely information from various external databases as they work. By doing so, RAG grounds the models' responses in reality, enabling them to produce more accurate and reliable text that is supported by actual data. This approach represents a useful step towards improving the trustworthiness and performance of LLMs, by ensuring that their outputs are based on information that can be fact-checked.

In this exploratory work, we apply a RAG-enhanced LLM to the scientific literature domain, specifically targeting the field of archives. This method leverages an architecture that dynamically provides the most relevant and contextually important proprietary, private, or dynamic data to the LLM during its operation. This integration enhances the model's accuracy and performance by grounding its responses on verifiable sources. Through a series of experiments, we demonstrate the model's improved ability to generate accurate and reliable text outputs, significantly reducing the incidence of hallucinations. Furthermore, we showcase the development of a website designed for the I Trust AI partnership to provide users with reliable LLM-generated content, emphasizing our commitment to fostering trust and reliability in AI-generated information.

Background

Large Language Models (LLMs)

LLMs are a major step forward in how computers understand and generate text, making them sound quite human-like. These models learn from large datasets, but it is often hard to tell how much of that information they can remember and use correctly. When LLMs create text, they do not have a way to check whether what they are saying is actually true. They do not work like search engines that show where their answers come from; instead, they create responses based on what a user asks them, and these responses are not always pulled directly from the data they were trained on. Sometimes, this leads to what we call "hallucinations," where the model comes up with information that is not true. It is important for users to remember this when using LLMs to make sure the information they get is reliable.

Hallucination

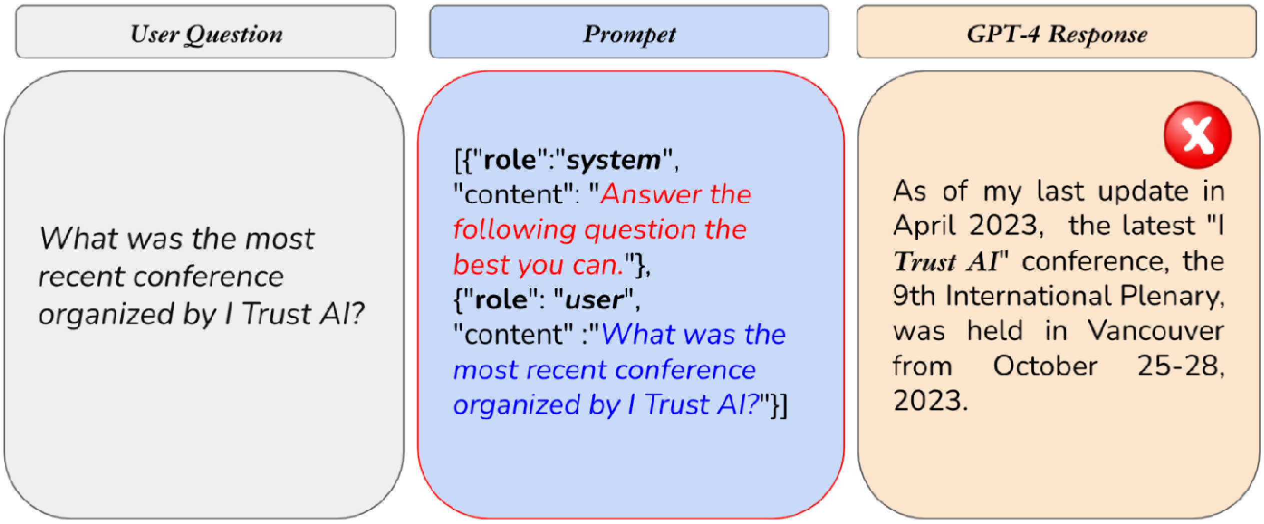
In the context of LLMs, “hallucination” refers to a phenomenon where the model generates text that is incorrect, or not real. Since LLMs are not databases or search engines, they would not cite where their response is based on. These models create text by building on the prompt a user gives them. Thus, the outcome is not always based on specific training data but is closely related to a user’s prompt. LLMs can exhibit hallucinations in various forms. Many studies have focused on understanding and mitigating them. A few examples of where LLMs can generate erroneous hallucinations are provided by Rawte et al. (2023):

a) *Sentence Contradiction*: This occurs when a model generates a statement that explicitly contradicts another statement in the same response or an earlier one.
b) *Prompt Contradiction*: This happens when the output contradicts the information in the instructions provided in the prompt.
c) *Factual Contradiction*: It is the generation of content that is factually incorrect or misrepresents established facts.
d) *Random LLM Hallucinations*: These are unpredictable and often bizarre statements or narratives that have no basis in the prompt or reality, showing a breakdown in the model's coherence.

Retrieval Augmented Generation (RAG)

The leverage of RAG in the context of NLP and LLMs was initially explored in two publications. Lewis et al. (2020) “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks” introduces the concept of RAG and Gao (2023) “Retrieval-Augmented Generation for Large Language Models: A Survey” builds on this idea and applies it to LLMs. Lewis et al. (2021) define RAG Models as follows: “models… which use the input sequence x to retrieve text documents z and use them as additional context when generating the target sequence y”. Lewis also describes RAG Models as being composed of two parts (i) *retriever* and (ii) *generator*. The (i) *retriever* would use a query (or prompt) to return a distribution over the text. The (ii) *generator* would take previous tokens to generate new tokens that would closely align with the distribution.

To understand this method, let’s study the user interaction with a LLM using RAG (see *Figure 1*) and one not using it (see *Figure 2*).



User Question

What was the most recent conference organized by I Trust AI?

Prompt

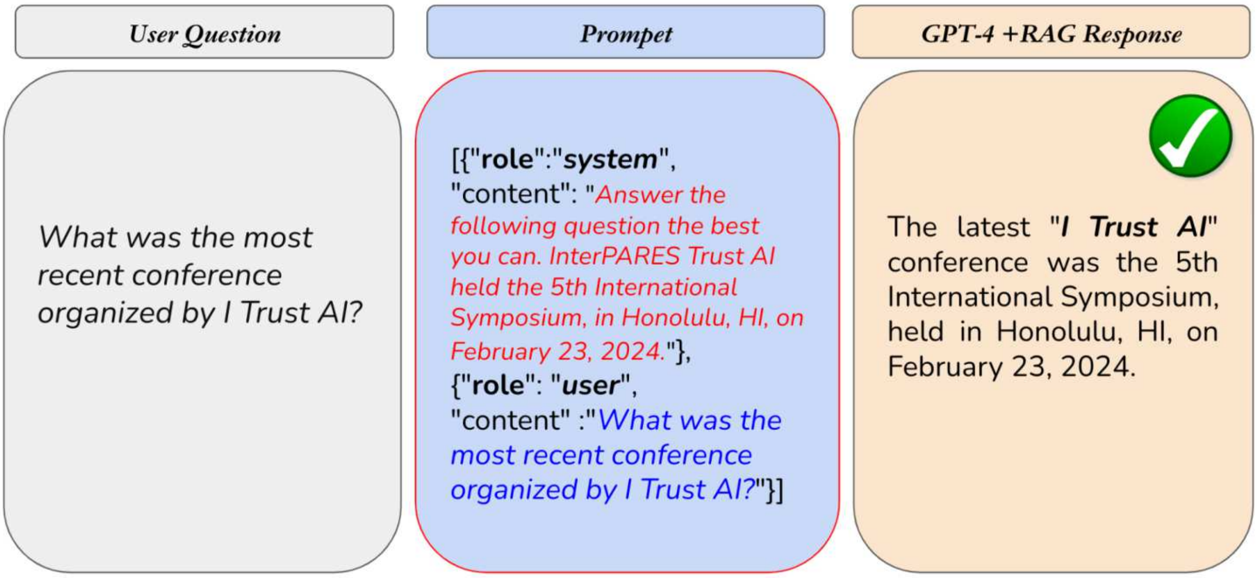
[{"role":"system", "content": "Answer the following question the best you can."}, {"role": "user", "content" :"What was the most recent conference organized by I Trust AI?"}]

GPT-4 Response

As of my last update in April 2023, the latest "I *Trust AI*" conference, the 9th International Plenary, was held in Vancouver from October 25-28, 2023.

Figure 1: GPT-4 incorrectly identifies the most recent *I Trust AI* conference as the 9th International Plenary, highlighting the limitation of not having real-time updates. All images by the authors.

LLM without RAG. *Figure 1* outlines a user's engagement with an LLM in a simple, four-step workflow that lacks RAG integration. It starts with a 'User Question', where the user poses a query. This question is then transformed into a 'Prompt', formulated to be understood by the LLM. Following this, the LLM generates a text-based 'Response', which may undergo 'Post Processing' to refine the content before being relayed back to the user. Without RAG, the model directly generates responses based on its training, which may lead to outdated or incorrect information, reflecting the system's limitations in ensuring the reliability and currency of the content it produces.



User Question

What was the most recent conference organized by I Trust AI?

Prompt

[{"role":"system", "content": "Answer the following question the best you can. InterPARES Trust AI held the 5th International Symposium, in Honolulu, HI, on February 23, 2024."}, {"role": "user", "content" :"What was the most recent conference organized by I Trust AI?"}]

GPT-4 +RAG Response

The latest "*I Trust AI*" conference was the 5th International Symposium, held in Honolulu, HI, on February 23, 2024.

Figure 2: Demonstrating the RAG model's capacity to deliver current and accurate information, demonstrating RAG's ability to provide current and verified information.

LLM with RAG. Incorporating Retrieval Augmented Generation (RAG) into a Large Language Model (LLM) significantly enhances the model's capability to provide up-to-date and accurate responses. The example in *Figure 2* demonstrates how a user's question about the most recent conference organized by I Trust AI is processed with the RAG-enhanced LLM. The system's prompt not only frames the user's query but also includes the most recent I Trust AI conference as well as the date it took occurred. This RAG-enabled prompt allows the LLM to deliver a current and correct response, highlighting RAG's vital role in ensuring that an LLM's output reflects the latest real-world developments.

Application: Archival Domain

Applying the principles of RAG to the archival domain involves a sophisticated model that seeks to enhance the accuracy and relevance of responses to queries about scholarly articles. For this application, the RAG model we propose utilizes 'Mixtral 8x7B' (Jiang et al., 2023), an open-source compound of eight billion-parameter expert models, ensuring a robust and diverse knowledge base. Additionally, the model incorporates 'FlagEmbedding' (Xiao et al., 2023), a state-of-the-art sentence embedding technique, to understand and generate contextually relevant prompts. For the archival data, we draw from a subset of the comprehensive datasets provided by the I Trust AI project. This rich source of information allows the RAG system to offer accurate and relevant answers, thereby significantly improving the user experience in scholarly research inquiries (see *Figure 3*).

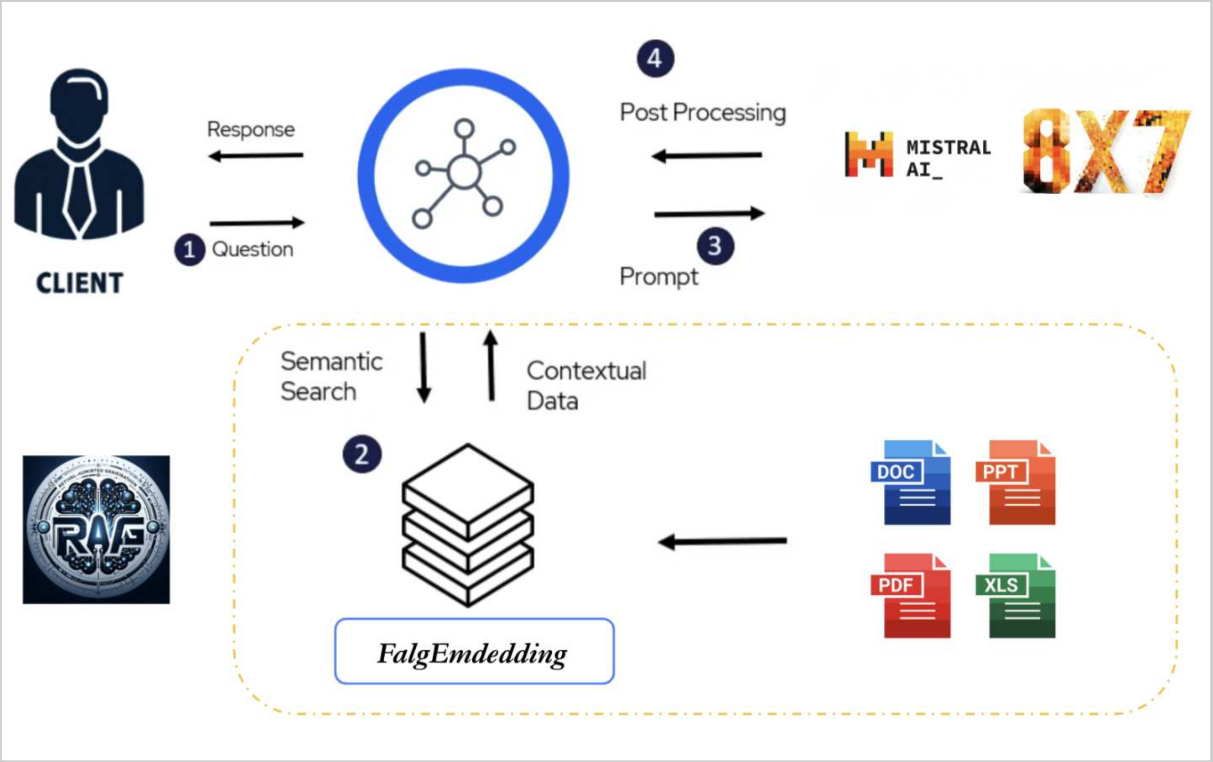
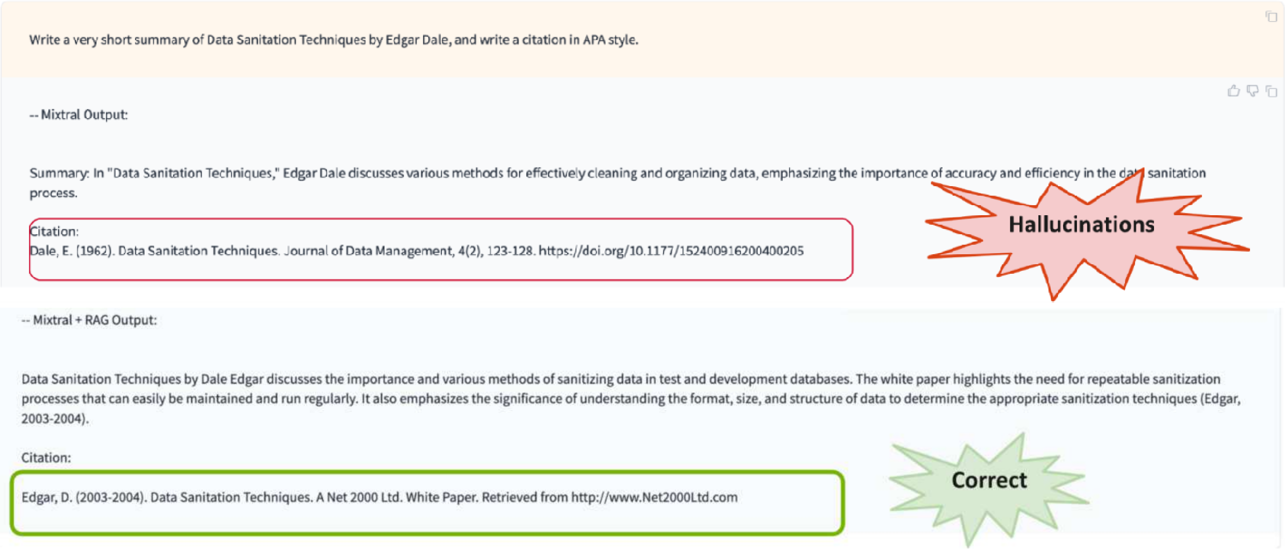


Figure 3: The RAG model flowchart depicts a client's question being refined by semantic search and contextual archive data, producing a well-informed response from the LLM.

Figure 4 shows that the integration of RAG with Mixtral 8x7B has a profound impact on enhancing the accuracy and reliability of generated content. By cross-referencing with updated databases, RAG allows LLMs to provide responses that are not only contextually rich but also factually correct. This is exemplified in the comparison where a non-RAG LLM output might yield outdated or "hallucinated" data, while a RAG-assisted LLM can reference current information, as shown in the example of accurately citing from a more recent source on Data Sanitation Techniques. This advancement in technology ensures that users receive the most relevant and accurate information available, marking a significant step forward in the practical application of AI in research and data analysis.



Write a very short summary of Data Sanitation Techniques by Edgar Dale, and write a citation in APA style.

-- Mixtral Output:

Summary: In "Data Sanitation Techniques," Edgar Dale discusses various methods for effectively cleaning and organizing data, emphasizing the importance of accuracy and efficiency in the data sanitation process.

Citation:
Dale, E. (1962). Data Sanitation Techniques. Journal of Data Management, 4(2), 123-128. https://doi.org/10.1177/152400916200400205

-- Mixtral + RAG Output:

Data Sanitation Techniques by Dale Edgar discusses the importance and various methods of sanitizing data in test and development databases. The white paper highlights the need for repeatable sanitization processes that can easily be maintained and run regularly. It also emphasizes the significance of understanding the format, size, and structure of data to determine the appropriate sanitization techniques (Edgar, 2003-2004).

Citation:
Edgar, D. (2003-2004). Data Sanitation Techniques. A Net 2000 Ltd. White Paper. Retrieved from http://www.Net2000Ltd.com

Figure 4: Comparison of outputs from Mixtral 8x7B with and without RAG, demonstrating RAG's effectiveness in providing current and accurate references, as seen in the precise citation of 'Data Sanitation Techniques'.

Conclusion

In this exploratory study, we addressed a critical issue of "hallucinated" responses – erroneous or fabricated content produced by LLMs. This integration into the archival domain, utilizing cutting-edge tools like 'Mixtral 8x7B' and 'FlagEmbedding', has shown that RAG can significantly enhance the reliability of LLMs, ensuring outputs are grounded in verified data. The successful application of RAG demonstrates its potential to transform the landscape of AI-driven research, offering a new level of accuracy and trust in automated content generation. With this technology, the I Trust AI platform stands as a testament to the possibilities of providing researchers and users with dependable, AI-generated information, thereby fostering a more trustworthy digital information environment.

References

Gao, Yunfan, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

Lewis, Patrick, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.

Rawte, Vipula, et al. The Troubling Emergence of Hallucination in Large Language Models--An Extensive Definition, Quantification, and Prescriptive Remediations. *arXiv preprint arXiv:2310.04988* (2023).

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., & Sayed, W. E. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Xiao, S., Liu, Z., Zhang, P., & Muennighof, N. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597* (2023).

Dr. ElMoatez Billah Nagoudi is a Postdoctoral fellow in the Deep Learning and NLP Group at the University of British Columbia focuseing on the development of large-scale language models for natural language understating and generation, neural detection of misinformation and machine-generated text, and multilingual machine translation. He completed a dual Ph.D. in Computer Science from Amar Telidji University of Laghouat, Algeria, in collaboration with Grenoble Informatics Laboratory, France. Before UBC he held multiple positions, including Assistant Professor at the CHL University, Algeria and a Member of the Computer Science and Mathematics Laboratory (LIM), Algeria (2012-2019).

Alcides Alcoba Inciarte is a research assistant at the University of British Columbia (UBC), where, for more than two years, he has engaged in projects focusing on natural language processing (NLP) and natural language understanding (NLU). With a Data Science Diploma from BrainStation and a Bachelor's degree in Applied Mathematics from SUNY Buffalo, he brings a robust foundation in data analysis, machine learning, and statistics to his work, and provides support to several studies in InterPARES Trust AI.

Dr. Abdul-Mageed is a Canada Research Chair in Natural Language Processing and Machine Learning, and Associate Professor in the School of Information and Department of Linguistics (Joint Appointment), and Computer Science (Associate Member), at The University of British Columbia.

UNESCO Audio Archives: AI for Metadata Enrichment

by Peter Sullivan and Eng Sengsavang

A diverse and multilingual audio heritage

From 2017 to 2020, the UNESCO Archives digitized 15,316 archival audio recordings dating from the 1950s to the 1980s - the majority of UNESCO's existing recordings on reel-to-reel magnetic tape.[4] The recordings document a remarkable range of topics, personalities, geographies, languages, and genres across four decades and in over 70 languages, including radio programmes, interviews, speeches, events, music, and more, reflecting the substance of UNESCO's work and its international, intergovernmental character. An additional 2,314 recordings from the same collection were repaired and digitized between 2021 and 2022. The digitized recordings - approximately 17,630 in total - now form a substantial part of UNESCO's digital archives.

Work is still ongoing to describe and make available the recordings for a general audience, so they can be studied, enjoyed, and widely used as rich sources of history. The central challenge facing UNESCO archivists is the description of the contents of the recordings. Scant information exists about the contents of many of them. While digitization of the recordings enables listening and rediscovery of their contents - and an opportunity to enhance the existing information about them - the current process of manually creating structured data for each recording is time-consuming and laborious. To date, about 1,000 recordings - less than 10% - are fully described and published.

Artificial intelligence offers opportunities to address some of the practical challenges related to describing and publishing the recordings. With this substantial dataset, it becomes possible to test how AI may help to both alleviate and enhance human work, and also to understand the challenges and potential biases of AI models - in short, to test how organizations can responsibly engage with AI. The particular characteristics of the digitized audio recordings - their historical, multilingual, and heterogeneous nature - will shape some of the challenges encountered when attempting to use machine learning interventions to describe them. These challenges, and the need to proceed within a conscientious framework, keeping in mind UNESCO's Recommendation on the Ethics of Artificial Intelligence, is guiding our work, and may find resonance among others around the world who care for sound archives, including the many sound archives on the International Memory of the World Register.

Metadata enrichment plan using AI models

Working with UNESCO Archives' research partner InterPARES Trust AI, which includes the University of British Columbia Deep Learning & Natural Language Processing Group, the driving question of our study can be summed as "how can AI enable better

[4] Digitized with support from the People of Japan, as part of the project *Digitizing Our Shared UNESCO History* (2017 to 2020).

description of archival audio recordings?" To address this question, we created a metadata enrichment plan that identified metadata elements of the audio description that are prime candidates for enrichment through AI models. These enriched metadata elements will enable more robust search of recordings by speaker and description, and identify for listeners which of over 70 languages is spoken in each recording - a task not possible even for staff working in a multilingual environment such as UNESCO. The plan also includes automatic generation of audio transcriptions and translations by different AI models, thereby establishing by comparison which model produces the most accurate results. The transcriptions and translations may eventually be made available alongside the audio recordings, but archivists will need to consider how to manage quality control of the transcriptions and translations before this happens.

The metadata enrichment plan includes investigating whether the use of a traditional archival method, diplomatic analysis, can be automatically applied to the audio recordings and their transcripts as a means of extracting key details from the structure of the recordings. While AI tools have become quite powerful in recent years, anchoring our approach in archival diplomatics theory and using knowledge about the underlying structure of various genres of audio recordings can potentially provide more control over the application of these AI tools. Given enough data, AI can almost certainly learn this structure automatically. However, the paucity of examples for any given style of recording, which would not allow for robust machine learning, makes a hybrid approach more appealing. In this approach, we use archival diplomatics methodology to identify patterns in the underlying structures of similar types of recordings, such as interviews and radio programmes, explicitly labelling the important structural parts, thus simplifying the AI problem to be solved.

Alongside this effort to merge traditional and modern strategies for enhancing the description of recordings, we also aim to use modern speech processing methods to identify the language and speakers on the recordings. Though some transcription methods are able to infer the type of language being spoken and directly transcribe it, language metadata may still be a valuable tool for researchers for finding material that may meet their search criteria. Most transcription tools rely on the knowledge of the language of a recording and, with over 70 languages included in the radio archives, having that knowledge is very important for the effective use of these tools. Similarly, while many recordings have important UNESCO personalities identified in the metadata, this is often an incomplete process, with many recordings having no information about who the speakers are, or providing only a partial list of the speakers. Speaker recognition tools offer potential answers on this front, allowing for the creation of an index of speakers automatically just from the vocal characteristics heard on the recordings.

Diplomatic analysis and AI for extractive summarization and audio speech recognition

Diplomatics emerged as a distinct science in the 17th century, aiming to understand and analyse the form of documents for the purpose of establishing their authenticity and to expose forgeries (Duranti, 1998, pp. 36-37). Still today, it has proven useful for contemporary archivists in the analysis of contemporary records.[5] For our purposes, it provides a gateway for extracting and labelling important structural components of the audio recordings, which can then be utilized alongside existing AI summarization techniques to describe them. The assumption is that straightforward: recordings in the same genre will have similar structural characteristics. For example, interviews and radio programmes explicitly provide descriptions of the nature of the recording, as a consistent element of their form, such as when an interviewer announces at the beginning of a recording 'who is being interviewed and about what', or when a radio programme host states the title of the programme and the subject or subtitle of that particular edition in the series.

The first step in the diplomatic analysis was therefore to identify the genres of recordings that may be candidates for this approach. These include: interviews; speeches; press conferences; and various types of radio programmes, notably educational or documentary radio programmes, reportage or commentary, and musical programmes. More genres or subgenres may be added as the analysis continues. For each genre, the multiple methods for analysis established in diplomatics are applied to several recordings in the same genre to identify any consistent or near-consistent structural properties of each genre, particularly at the beginning, middle, or end portions of the recordings. In addition to the digitized recordings, transcripts of the recordings are used in the analysis. These can be generated in .TXT format using OpenAI's Whisper (Radford et al., 2023) with some minor Python scripting, with timestamps at 1- to 6-second intervals. The transcripts enable quick scanning of the form and structure of any given recording, thereby acting as useful tools of analysis.

This process of documentary criticism has resulted in several observations that may potentially be applied as labels for AI summarization and automatic description techniques. For instance, radio programmes often begin with musical jingles of between 10 to 30 seconds, often signalling the beginning of a scripted and structured educational or documentary radio programme on a single topic or on multiple topics. Interviews, speeches, and press conferences often begin with an introduction to an individual by a narrator or host who invokes the name of the main speaker. In speeches and press conferences, the main speaker often introduces the general topic at hand within the first few minutes of their speech. As this work continues, we hope to produce a usable set of labels to support accurate AI summarization and automatic description of many of the recordings.

[5] Luciana Duranti was the first to introduce it to contemporary archivists and is today the foremost expert on diplomatics in the archival field.

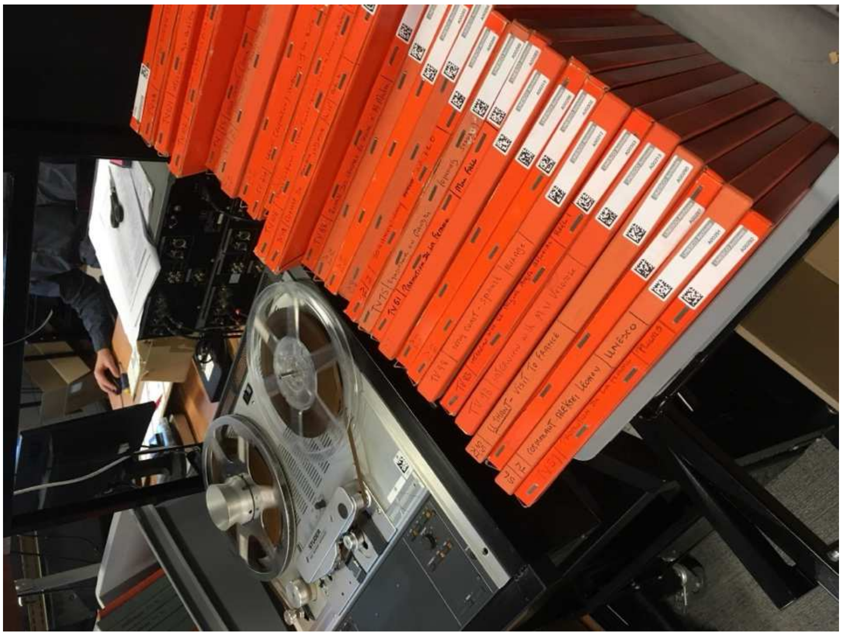


Figure 1: Digitizing UNESCO audio recordings. Credit: Adam Cowling.

Challenges of multilingual audio and AI

Initial work with the archives looked at the quality of automatic speech recognition transcripts and accuracy of language ID predictions. Our early findings using Whisper indicated that, for high resource languages such as French and English, the transcriptions were quite reasonable. However, this was only the case when the language ID prediction was accurate.

An additional challenge for language ID present in the archives is the number of multilingual speakers. To give an example, we have many recordings of Vittorino Veronese, the UNESCO General Director from 1958 to 1961, including recordings in English, German, French, Spanish, as well as his mother tongue, Italian. Multilingual speech goes hand in hand with a wide variety of accents and language backgrounds (Sullivan et al., 2022). But accents can prove particularly problematic for language ID tools and, to verify whether this would be an issue in our case, we examined the performance of several off-the-shelf language ID tools on a set of second language English speakers from the EuroParl (Wang et al., 2021), as well as a set of multilingual speakers in the radio archives. While Whisper appeared to do fairly well on the accented English speech (with Large V3 performing at 94% accuracy), we found the language ID capabilities of Meta's MMS6 fell short (only recognizing 11% of the utterances as English). On the multilingual speakers set, we found that the newest version of Whisper (Large V3) performs at 92% accuracy, which shows promise for addressing our language metadata problems.

To tackle our speaker indexing problem, we have also looked at the robustness of speaker recognition tools. These tools are often based on extracting a representation of the speaker's voice from the recording, most recently either through deep neural network based x-vectors (Pratap, 2023) or CNN based (Despanques et al., 2020). These representations ideally only capture the important characteristics of someone's voice, but unfortunately may also capture linguistic, age, and gender bias (Huitri and Ding, 2022). Cross-age shifts in speaker vocal characteristics are an open problem for speaker recognition (Qin at al., 2022), and are especially impactful in archival recordings, which may include the same speakers over many years (in the radio archives, we have a few such speakers over 20 years). Similar to cross-age issues, as speaker representations may capture linguistic information, are phenomena like code-switching; or, simply speaking a different language on a different recording may have an impact on recognition accuracy. Our work so far has confirmed that both the cross-age and cross-lingual nature of the recordings impact the similarity of speaker representations, and our work continues to develop tools that will help overcome this gap.

Figure 2: UNESCO audio recordings. Credit: Maeva Nguyen.

Future steps

Our work has demonstrated the importance of robust AI when working with archival audio materials, and we look forward to future work in improving the customization of the existing methods to support archival needs. Our next steps include building a benchmark for assessing speaker indexing, particularly focusing on the impact of multilingual and cross-age speakers. Additionally, our work in combining diplomatic analysis and automatic summarization continues, with our next step being the finalization of a set of annotated transcripts covering various genres that can be used for summarization experiments.

References

Desplanques, B., Thienpondt, J., Demuynck, K. (2020) ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proceedings Interspeech 2020*, 3830-3834, doi: 10.21437/Interspeech.2020-2650

Duranti, L. (1998). *Diplomatics: New Uses for an Old Science*. Society of American Archivists and Association of Canadian Archivists in association with The Scarecrow Press, Inc., Maryland.

Hutiri, W. T., & Ding, A. Y. (2022, June). Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 230-247).

Pratap, V., Tjandra, A., Shi, B., Kundu, P. T. A. B. S., Elkahky, A., Fazel, Z. N. A. V. M., Auli, M. (2023). Scaling Speech Technology to 1,000+ Languages. *arXiv preprint arXiv:2305.13516*.

Qin, X., Li, N., Chao, W., Su, D., Li, M. (2022) Cross-Age Speaker Verification: Learning Age-Invariant Speaker Embeddings. In *Proceedings Interspeech 2022*, 1436-1440, doi: 10.21437/Interspeech.2022-648

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5329-5333). IEEE.

Sullivan, P., Shibano, T., & Abdul-Mageed, M. (2022). Improving automatic speech recognition for non-native English with transfer learning and language model decoding. In *Analysis and Application of Natural Language and Speech Processing* (pp. 21-44). Cham: Springer International Publishing.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., ... & Dupoux, E. (2021, August). VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 993-1003).

Peter Sullivan is a PhD student at the University of British Columbia's School of Information.

Eng Sengsavang is Reference Archivist at UNESCO Archives. She is co-editor with Jens Boel of Recordkeeping in International Organizations: Archives in Transition in Digital, Networked Environments *(Routledge, 2021).*

Datafying Archives for Privacy Protection

by Alcides Alcoba, Paige Hohmann and Jim Suderman

There are good reasons to consider applying Artificial Intelligence (AI) to protecting information privacy. According to the United Nations Conference on Trade and Development, 80% of nations already have privacy protection laws either in force (71%) or in draft (9%).[7] Privacy protections are also long-term, often extending beyond the operational lifespan of records. This results in inactive records being transferred from a creating office to archivists or other record custodians with only cursory descriptions and minimal information regarding personal information (PI). Yet the responsibility for protecting privacy remains unreduced and so archivists receiving large transfers of minimally described records often have little option other than to simply close the records.[8] To address some of these issues, a study in the InterPARES Trust AI (I Trust AI) project is exploring the development of an AI model that predicts the existence of all types of personal information as a means to help archivists manage privacy responsibilities relating to records in their care more effectively.

Applying AI to protect personal privacy is not a new idea. From the literature, it appears that existing tools primarily address personal information that takes the form of regular expressions. For example, nine of the twelve global PI entities Microsoft Presidio is trained to detect are dependent on pattern matching, with just three dependent on "custom logic and context."[9] Similarly, of the twenty-two labels in Amazon Comprehend categorizing PI, at least twelve are numbers or dates with another five (pin, email, driver ID, aws access key, aws secret key) having reasonably well-defined syntax.[10] Based on published information, it does not appear that AI-enabled PI prediction tools have evolved to integrate some of the remarkable breakthroughs based on contextual information of recent years. For example, the findings of Jason R. Baron, Mahmoud F. Sayed, and Douglas W. Oard regarding an AI model trained to predict whether reviewed text might be governed by the deliberative process privilege.[11] The apparent lag in

[7] UNCTAD, "Data Protection and Privacy Legislation Worldwide," visited 12 March 2024. For the purposes of this article the terms Personal Information (PI), Personally Identifiable Information (PII), and Personal Data (PD) are considered synonymous (see applicable Glossary entries of the International Association of Privacy Professional).

[8] Legislated privacy protection *never* expires in British Columbia. Instead, organizations must "destroy documents containing personal information or make the information anonymous as soon as it is reasonable [to do so)…" Office of the Information & Privacy Commissioner, British Columbia, "A Guide to B.C.'s Personal Information Protection Act for Business and Organizations" (2015), p. 38.

[9] The three are NRP (nationality, religious or political group), Location (politically or geographically defined), and Person (full name). See Microsoft Presidio, "PII entities supported by Presidio," visited 12 March 2024.

[10] The exceptions are Name, Address, Age, Username, Password. See AWS, "EntityLabel," visited 12 March 2024.

[11] While PI is not a factor in assessing the applicability of the deliberative process privilege, the study illustrates the potential of AI to process more complex, context-driven factors. Jason R. Baron, Mahmoud F. Sayed, and Douglas W. Oard. Providing More Efficient Access to Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege. *J. Comput. Cult. Herit.* 15, 1, Article 5 (January 2022), https://doi.org/10.1145/3481045.

integrating these more advanced approaches into commercially available applications may be due to i) uncertainty regarding the robustness and interpretability of deep learning approaches;[12] ii) the impossibility of establishing a comprehensive typology of personal information; and iii) the scarcity of large bodies of heterogeneous record types containing a wide range of simple and complex PI suitable for training AI models.

The I Trust AI study starts from the assumption that some of the limitations noted above might be addressed if the development of the model conformed to the juridical and organizational contexts where it will be used. To that end, the juridical context of the study was set as Canada with the organizational context of the University of British Columbia (UBC). The model will be trained on data extracted from a highly sensitive *corpus* of administrative records from UBC - a process that itself entails considerable effort.

Two pathways toward securing this fit-for-purpose training substrate appeared when we attempted to convert our largely static image-based[13] *corpus* into a dataset: one of them is administrative while the other is conceptual. In the first instance, the records identified and ultimately obtained for analysis are compiled dockets prepared in response to Freedom of Information (FOI) requests received by the UBC Office of University Counsel (OUC). The contents are certainly heterogeneous given that the dockets are composed of records created across numerous departments, functions, and formats, and are unified only by their correspondence to the topic or subject of the FOI request. Normal OUC procedure dictates that prior to records release to the applicant, the dockets are evaluated and severed by a qualified human to ensure that no inappropriate or unlawful disclosure of third party PI takes place.

Because the *corpus* features both severed and non-severed versions of the dockets, repurposing it for research presents a relatively high level of risk to UBC.[14] The first risk is legal and reputational risk. As a public body, UBC is subject to British Columbia's *Freedom of Information and Protection of Privacy Act.*[15] The second risk is potential harm to human subjects as an outcome of secondary use of the records for research. In both cases, risk is reduced to a tolerable level by demonstrating the ability to keep the data secure. This capacity is documented in two reciprocal review processes leading to approvals to release the *corpus* and proceed: i) with the Data Governance Committee,[16] operated by the UBC Office of the CIO (Chief Information Officer), and ii) with the Behavioural Research Ethics Board operated by the UBC Office of the VP, Research and

[12] Gasparetto, A.; Marcuzzo, M.; Zangari, A.; Albarelli, A. A Survey on Text Classification Algorithms: From Text to Predictions. *Information* 2022, 13, 83. https://doi.org/10.3390/info13020083.

[13] Consisting of PDF files compiled from various UBC offices and a variety of formats.

[14] The sensitivity profile of the records *corpus* is characterized by Section 22 of the British Columbia Freedom of Information and Protection of Privacy Act (BC FOIPOP) https://www.bclaws.gov.bc.ca/civix/document/id/complete/statreg/96165_00

[15] British Columbia, Freedom of Information and Protection of Privacy Act.

[16] UBC, Office of the Chief Information Officer, Data Governance Services. Access UBC Data. https://cio.ubc.ca/data-governance/data-governance-services/access-ubc-data.

Innovation,[17] in compliance with the Tri-Council Policy Statement Guidelines.[18] At the conclusion of these processes, this pathway culminated in a "binder" of documents, including a research ethics approval certificate, a legal letter of approval and data release, a research data management plan, a formal agreement between the University and the Principal Investigator, and a procedural framework including requirements for training and non-disclosure agreements for the research team.

The second (conceptual) pathway follows the research team's exploration of how best to teach a machine to "read" the documentary context that is largely absent from current, content-focused privacy detection models. In this case, what is needed is an automated "reading" aimed at replicating the processes of reviewing and severing documents of their embedded PI/PII extending beyond simple regular expressions. This revised and enhanced approach should account for the persons, acts, and conventional structures and components of modern administrative records.

Archival diplomatics was adopted as a conceptual framework for correlating document characteristics to the probability of presence of PI/PII. As a method, a diplomatic analysis is applied at the item level, and describes and categorizes key documentary features that are the unavoidable outcomes of the actions that make record creation necessary in the first place. The records in the initial training sample are removed from their original contexts of creation; to bridge this gap, the UBC institutional records retention schedule was used to classify the records in terms of their function, prescribed disposition and archival value, and predicted presence of PI risk.[19]

With the *corpus* now accessible to research activities through pathway one (administrative) and semantically clarified through pathway two (conceptual - diplomatic) the challenges now bridge to the technical realm through a set of functional requirements for labelling a training set of records.

Labelling should inform the training of a model that can identify:

1. The beginning and end of each constituent record within a consolidated PDF;
2. The form of each document; and
3. The specific components corresponding to intrinsic diplomatic elements.[20]

As we are working with private and sensitive data, steps had to be taken to ensure that it is well protected and kept within the UBC technology infrastructure. We also needed a space where our Data Science researchers can experiment, and where our archivists can visualize and label the *corpus*. We requested permission to use Sockeye to maintain the

[17] UBC, Office of Research Ethics. Behavioral Research Ethics. https://ethics.research.ubc.ca/behavioural-research-ethics

[18] Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, TriCouncil Policy Statement: Ethical Conduct for Research Involving Humans, December 2022. https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html.

[19] UBC, Records Management Office. Retention Schedules. https://recordsmanagement.ubc.ca/schedules/

[20] Intrinsic diplomatic elements are observable components of the document's intellectual articulation that can be understood independently of the document's semantic content. Examples of intrinsic elements in the first iteration of labelling include protocol, text, and escatochol, see: Rogers, Corinne, "Diplomatics," in Luciana Duranti and Patricia Franks, eds., *Encyclopedia of Archival Science*, Rowman & Littlefield Publishing Group. Lanham, 2015, p. 178.

data and build our workspace. Sockeye is the UBC's Advanced Research Computing facility[21] that, with its wealth of resources and ease of access, allowed us to quickly prepare it as a sandbox that satisfies the study's requirements, which means that

1. Data is stored safely;
2. Only the members of the study have access;
3. Its infrastructure allows us to install and request software; and
4. Strong documentation and support are provided.

We prepared a workspace ecosystem following Sockeye's documentation, installing the latest versions of Python, and identifying JupyterLabs as an Integrated Development Environment (IDE) for interfacing with the PDFs and writing code. We also installed Label Studio,[22] a labeling interface that runs on Python that can directly read the PDFs from Sockeye. We prepared a guide so that every user is able to use all these tools.

It is important that we understand the properties of data that machine-learning models require so that we, later on, can define our labeling instructions. We brainstormed several Natural Language Processing (NLP) and Computer Vision techniques, envisioning the necessary steps that had to be taken to go from PDF files to annotated text. We discussed an initial data pipeline as follows:

1. Use Object Detection/Segmentation to draw bounding boxes around text.
2. Convert the generated bounding box images to text via Optical Character Recognition.
3. Use Name Entity Recognition to detect persons and organizations.
4. Perform other NLP tasks such as Question Answering (QA), Summarization, and Text Classification to extract other important information relating to Diplomatic Elements and Image Classification to identify different parts of a document:
    a. QA to ask questions such as "Who is the author?", "What is the intent ?", "When was it written?"
    b. Summarization to help archivists get a quick idea of the document, which may be hundreds of pages long.
    c. Image/Text Classification to indicate the protocol, eschatocol, the start and end of each document, the kind of document, etc.

Furthermore, we discussed the idea of using Large Language Models (LLMs) to directly process the *corpus*. However, we may have to limit the amount of computing power used at runtime, as we may be building a tool for archivists who may not have the resources to run powerful models.

For the set-up of the Label Studio labeling interface, we display one page of each document at a time. There are instructions for drawing bounding boxes, with labels "Protocol", "Eschatocol", and "Body", and tags for classifying the document as "Letter", "Email", or "Agreement". The output will be a JSON file with page-level annotations for each file.

[21] UBC, Advanced Research Computing. UBC ARC Sockeye. https://arc.ubc.ca/compute-storage/ubc-arc-sockeye

[22] Tkachenko, Maxim, Malyuk, Mikhail, Holmanyuk, Andrey, & Liubimov, Nikolai. (2020). Label Studio: Data labeling software. Retrieved from https://github.com/heartexlabs/label-studio

This article describes a study in progress which is shared here because applying concepts from archival diplomatics and external context to consider documents and their meaning is new to the use of AI-enabled tools and techniques.The challenge of establishing an appropriately labelled dataset for training such a model has been the focus of the study to date. The value of the study, as perceived by its participants, is a far richer, more reliable assessment of document contents for a variety of purposes beyond privacy protection.

Alcides Alcoba is a Research Associate at the Deep Learning NLP Group, University of British Columbia (UBC), Vancouver, Canada.

Paige Hohmann is the archivist at UBC's Okanagan Campus and an ITrust AI researcher.

Jim Suderman is the former Director of the City of Toronto's archives, records management, and FOI programs and an ITrust AI researcher.

Balancing Act: Navigating the Nexus of AI, Privacy, and Accessibility in Archives

by Victoria L. Lemieux

Introduction: The AI Privacy Problem

Globally, critics are sounding the alarm about the way that AI, particularly large language models, are disclosing personal information from information gathered from a wide range of sources, including public archives (see, e.g., Das et al., 2023). As an increasing number of public archival institutions are digitizing their materials and putting them online, or acquiring contemporary born digital records, the risk of exposing individuals' personal information when providing public access to archives has risen exponentially. Personal data leaks from AI can take many forms, from accidental disclosure by archival staff to data gained by deliberate attempts to by-pass privacy and security controls in institutions often ill equipped to defend against such attacks.

As archivists have a mission both to provide access to records and protect personal information, archivists now must comply with a growing body of laws, such as exemption provisions found in freedom of information laws or provisions of data protection legislation, that have arisen as a response to widening concerns about privacy and confidentiality in the digital age. Interpretation and application of these laws adds complexity to the archivists' task of providing access to archival holdings.

This article provides an overview of two approaches that archival researchers are exploring to use AI to address the threat to privacy presented by it, in essence fighting "fire with fire". The first approach seeks to automate the "sensitivity reviews" that archivists undertake in many jurisdictions before they can provide public access to materials that could contain personal information. This approach typically uses Natural Language Processing (NLP) and AI to identify personal information within an archival corpus so that any sensitive information can be "masked" before public release. Work by an international team of researchers with InterPARES Trust AI (I Trust AI) seeks to leverage archival science theory to improve the previously mixed results of this approach. The second approach involves treating all materials held in archival corpora as confidential and uses novel Privacy Enhancing Technologies to defend against accidental or deliberate data leakage. This approach seeks to enable researchers to extract valuable information from archival documents without providing direct access to them, thereby eliminating the risk of privacy breaches.

Using AI to Automate Archival Sensitivity Reviews

Archivists are increasingly using automated solutions to conduct sensitivity reviews of archival documents (McDonald et al., 2019; Baron et al., 2022). These solutions focus on identifying, anonymizing, or redacting personal information found in documents. Early techniques used regular expressions for text searching, but this approach has proven costly and inefficient, particularly as it struggles with changing contexts.

To overcome the limitations of early techniques, archival researchers have been experimenting with NLP tools to identify personal information in texts. Various NLP strategies have been applied, such as using text classification to identify document categories containing personal information (McDonald et al., 2019). However, these methods still require significant human involvement for training and prescreening.

Recent experiments have also tested AI for text classification in sensitivity reviews, though not yet with great success. Approaches like term frequency-inverse document frequency (TF-IDF) combined with Support Vector Machine (SVM) classifiers and neural network architectures like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) have been compared, but accuracy remains relatively low (Franks, 2022). These methods often produce many false positives and negatives and struggle with complex privacy contexts, such as when non-sensitive information combines in a way that discloses sensitive personal information (referring to the "mosaic effect").

Additionally, AI models used in these tasks face the challenge of potentially leaking personal information themselves. Model properties can be exploited in attacks, revealing sensitive personal information used in model training (Lemieux and Werner, 2024). Overall, while there is ongoing research into automated tools for managing sensitivity reviews of archival documents, these techniques are still evolving and face significant challenges.

Making it Possible to Provide Access *without* Providing Access: Using Privacy Enhancing Technologies

Since automating sensitivity reviews remains challenging and archives increasingly hold sensitive personal information, I Trust AI researchers also are experimenting with an emerging class of technologies known as Privacy Enhancing Technologies (Lemieux and Werner, 2024). These technologies – which include: homomorphic encryption; Trusted Execution Environments; Secure Multiparty Computation; Differential Privacy; Personal Data Stores; Privacy Preserving Machine Learning; and Synthetic Data (Royal Society, 2019) – allow for AI-enabled analysis of archival documents without requiring that researchers gain direct access to the documents.

One novel approach that I Trust AI researchers are working on will combine a technique increasingly used by researchers in the Digital Humanities called Distant Reading (Moretti, 2013) with Decentralized Privacy Preserving Federated Machine Learning to protect personal information from exposure. Distant Reading uses Text Mining, Natural Language Processing, and AI to help researchers analyze large archival corpora. Typically, the output of such analyses is a visualization that represents broad patterns that can be gleaned from archival documents, such as the communication patterns between geolocations, public sentiment over time, or topics or themes represented in a corpus of archival text. Using Distant Reading enables researchers to learn from large archival corpora without having to inspect and analyze each individual document, which, in turn, relieves archivists of the burden of undertaking sensitivity reviews before providing public access to their holdings.

Distant Reading alone will not prevent data leakage, however, unless the algorithms used are privacy preserving. This calls for a solution that combines Distant Reading with the use of Privacy Preserving Federated Machine Learning - a collaborative learning method wherein multiple parties train a model without centralizing their data or exposing it to other parties (Chen et al., 2021). The first Federated Learning framework was introduced by Google in 2016 to build Machine Learning models using data across multiple devices. Newer approaches to Federated Machine Learning improve upon earlier techniques and provide protection against inadvertent data leakage by machine learning models. In Decentralized Privacy Preserving Federated Machine Learning, for example, multiple computers, possibly located across different organizations or geographies, collaboratively train a machine learning model while keeping their data on-premise to preserve privacy as in classic Federated Machine Learning, but also require that each computer performs computations on its own dataset and then shares only limited insights, like updated weights or gradients, with other computers (Li et al., 2022). These updates can be shared in a way that protects the privacy of the underlying data, using privacy preserving techniques to prevent a single computer from revealing too much information or having too much influence over the model. Decentralized Privacy Preserving Federated Machine Learning also has the advantage of protecting models against attacks on their integrity and single points of failure.

The application of Privacy Enhancing Technologies to help make it safe to apply AI in the analysis of archival holdings is still in its infancy and therefore has many limitations and unknowns, but I Trust AI researchers are hopeful that the techniques they are experimenting with may help with a difficult balancing act as archivists juggle AI, privacy and accessibility in archives.

References

Baron, J. R., Sayed, M. F., & Oard, D. W. (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. *ACM Journal on Computing and Cultural Heritage* (JOCCH), 15(1), 1-19. https://doi.org/10.1145/3481045

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., & Liu, Y. (2021). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 54(4), 1-40. https://doi.org/10.1145/3447744

Das, S., Lee, H-P, Forlizzi, J. (2023). Privacy in the Age of AI. *Communications of the ACM*, 66(11). https://dl.acm.org/doi/10.1145/3625254

Jason Franks (2022). Text classification for records management. *Journal on Computing and Cultural Heritage (JOCCH)* 15(3),1–19. https://doi.org/10.1145/3485846

Lemieux, V. L., Werner, J. (2024). Protecting Privacy in Digital Records: The Potential of Privacy-Enhancing Technologies. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 16(4), 1-18. https://doi.org/10.1145/3633477

Li, D., Han, D., Weng, T. H., Zheng, Z., Li, H., Liu, H., ... & Li, K. C. (2022). Blockchain for federated learning toward secure distributed machine learning systems: a systemic survey. *Soft Computing*, 26(9), 4423-4440. https://doi.org/10.1007/s00500-021-06496-5

McDonald, G., Macdonald, C., Ounis, I. (2019, March). How sensitivity classification effectiveness impacts reviewers in technology-assisted sensitivity review. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 337-341).

Moretti, F. (2013). *Distant reading*. New York, NY: Verso Books.

Royal Society (2019). *Protecting privacy in practice: The current use, development and limits of Privacy Enhancing Technologies in data analysis*. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Protecting-privacy-in-practice.pdf

Victoria Lemieux currently holds a position as Professor of Archival Science at the University of British Columbia's School of Information, an affiliated faculty position in UBC's Department of Electrical and Computer Engineering, Faculty of Applied Science, and is a member of UBC's Institute for Computing, Information and Cognitive Systems. She is also founder and co-lead of Blockchain@UBC, the University of British Columbia's multidisciplinary blockchain research cluster. Her academic research focuses on risk to the availability of trustworthy records and how these risks impact upon transparency, financial stability, public accountability, and human rights.

The Role of AI in Identifying or Reconstituting Archival Aggregations of Digital Records and Enriching Metadata Schemas

by Stefano Allegrezza, Maria Mata Caravaca, Massimiliano Grandi, Mariella Guercio, Bruna La Sorda

Using AI to build or recreate archival aggregations: The survey and its results[23]

The overall goal of the I Trust AI study entitled "The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas"[24] is to investigate the ability of Artificial Intelligence to support creation (or recreation) of archival aggregations in order to address the issue of non-aggregated, unarranged, or de-contextualized records (both in the current and semi-current phases of their lifecycle). In many public administrations and private companies, documents are neither classified nor aggregated. In other cases, aggregations of documents exist but are not appropriate, and this results in an uncontrolled number of documents that are unsorted, misplaced, and hard to find. In many cases metadata elements - necessary to ensure the reliability, trustworthiness, quality and sustainability of archival appraisal and acquisition - are missing. Despite progress in the development of technology supporting records management, current software applications are not very helpful in carrying out the activities of identification and maintenance of the original documentary relations necessary to ensure the evidentiary capacity of the records, thereby providing a qualified basis for the creator's accountability.

The research question that this study aims to answer (can AI tools help build or recreate archival aggregations and generate metadata schemas for them?) is crucial and complex to answer. The main difficulty concerns the ability to evaluate AI technologies, assess the risks associated with their use, understand their potential, and develop effective methods for collaboration between archivists and AI specialists in this specific domain. In particular, the study has been planned with the aim of supporting the archival understanding and knowledge of this area by investigating and analysing the most known and promising AI solutions in the specific field of archives and records management and, more specifically, in the creation and/or discovery of the archival relationships amomg the records, and their original aggregations.

This goal implies a comprehensive approach, based first on the review of the main platforms available, which were selected on the following two main criteria: a clear statement that document/record management is one of the objectives of the AI based application, and an expression of interest for the archival dimension, either explicitly stated or easily understandable on the company website. The effort made to limit the number of solutions to be taken into consideration and to restrict the survey to sellers

[23] The full report was published on November 1st 2023, and is available here: https://interparestrustai.org/assets/public/dissemination/Report-CU05-Survey-of-the-Companies_v121.pdf.

[24] CU05, https://interparestrustai.org/trust/about_research/studies.

whose products are relevant to archives and records management has required time, and may have not been free from errors and misinterpretations. Other parameters adopted for the selection of sellers have included the analysis of the specific portfolio of the companies, their involvement in the domain of archives and records, and their compliance with relevant regulatory frameworks and standards.

Twenty-eight companies corresponding to the above parameters were identified and invited to take part in the survey and answer a very detailed questionnaire. The questionnaire was designed for the systematic collection of the information necessary to an adequate assessment of the applications intended to support the reconstitution of archival aggregations, as well as metadata enrichment. The questionnaire included the description of the achievements of the companies, their compliance with the archival regulatory framework, the specific capabilities of the solutions for recordkeeping, the analysis of the AI technologies used, and the identification of key performance indicators. Thirteen companies completed the questionnaire and accepted to be contacted (Figure 1).

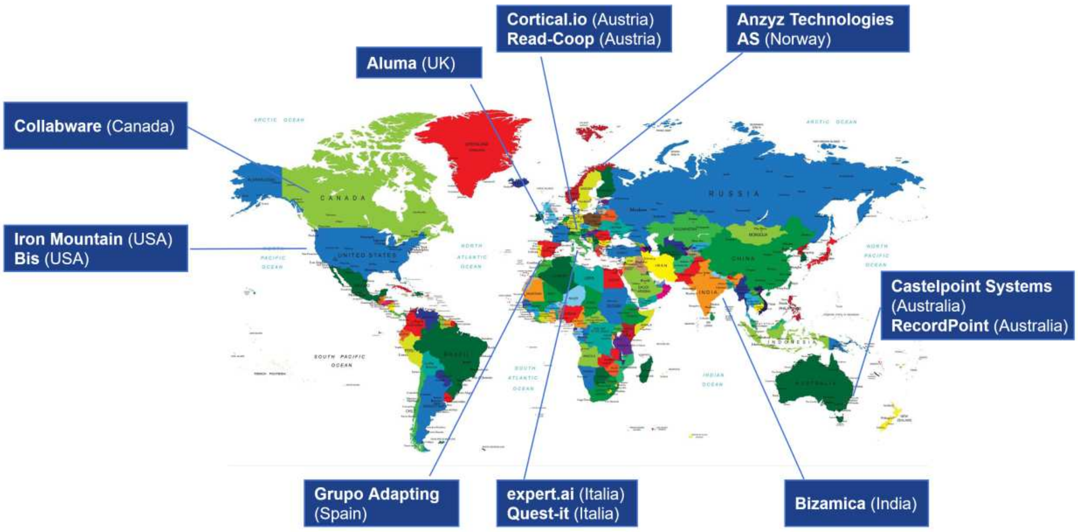


Figure 1. The 13 companies that answered the questionnaire (image by authors)

All the companies interviewed have developed solutions based on artificial intelligence technologies to manage, index, and classify structured, semi-structured and unstructured data through automatic learning techniques and automatic data extraction. More specifically, the automatic classification of records, which uses the support of a classification scheme, is offered by almost all the companies that replied to the questionnaires (10 out 13), including those not specifically involved with archives and records management. Among the companies whose platforms do not feature automatic classification, one mainly deals with handwritten text recognition and the other two focus on indexing and extraction of metadata elements. Some of these stated that their applications could be modelled or trained to categorise records.

Three companies familiar with the principles of archives and records management replied that their platforms analyse the metadata elements available both in the records and in the aggregations of which they are part. Six other companies concisely answered that their platforms can categorise documents according to a records classification scheme. One company declared that its platform can be trained to recognize document types and apply any kind of classification scheme.

With reference to the creation of archival aggregations (files and series), 10 out of the 13 companies that were interviewed stated that their applications are able to file records in their related folders, case-files or groups in an automatic or semi-automatic way, although with limitations. In relation to additional questions in this area, such as the possibility that applications make inferences about which records belong or might belong to the same group or business process (e.g. same case-file, subject-file, series, fonds), answers were positive in the range of 9 out 13, indicating that inferences are made as follows:

- "Based on content and/or context";
- "Case-folder might be found using AI models in a trained system";
- "Only with user-created content rules to specify the categories and requirements for categorization. Once these content rules are created and enabled, the system can make these decisions on newly added records";
- "If there is metadata to represent those processes (e.g. a case file number)";
- "The tools can output aggregated extractions exposing specific relations between extracted entities".

When companies were asked if their application was able to make inferences about the organisation or person that filed the records, even when relevant metadata elements for their identification were missing, the number of positive answers decreased to 7 of 13. Those responding in a negative way, made the point that inference could not be made directly without metadata.

In summary, developers of applications featuring automatic classification capabilities asserted that applications are also able to file records or logically link records to their related aggregation, as well as make inferences to achieve this goal. They also stated that this automation may apply classification or labelling schemes, and it is based on how the applications work or how they are trained, that is, according to record type, case-file, metadata, record content and context. These assertions are perhaps too optimistic. Both classification and filing/aggregation have different connotations depending on the archival tradition or records management standards. These may distinguish between classification and aggregation, or not, or may mix them up. This is, in some way, reflected in the answers to the questionnaire, where is not always easy to understand if applications categorise and/or set aside records based on established classes and/or files, and to what extent they proceed following a records classification scheme and/or a file plan, or just an established list of labels or tags.

With reference to the capacity of the applications of re-constituting archival aggregations that had lost their order, only five respondents out 13 answered positively and explained that they basically extract data from records content or from metadata (including the classification codes) to propose aggregations or relations among records. The other eight companies replied negatively, adding that reconstitution of lost order can be done only manually, or that they could "provide restored records to users in the case of data loss at the content source, but cannot reconstitute aggregations". The interviewed companies were also asked whether their applications were able to index records in order to provide information about related links or aggregations among records. Six out of 13 answered positively, while the other companies replied in an undetermined or negative way. Thus, according to the survey results, reconstituting records aggregations is not the main focus of companies developing AI technologies: it is a difficult task when contextual data are lost and it seems to require a more evolved technology.

In contrast, all the respondent companies have developed solutions to extract metadata, by using for example parsers to extract required entities or external automatic OCR applications to fill the metadata fields of the document to describe it or generate metadata elements from the content, where OCR is possible. However, some companies noticed that the quality of the text extracted through OCR processing is questionable and expensive to run.

The development of capabilities concerning appraisal and the implementation of retention schedules has not been so far a specific goal of the companies that have taken part in the survey, as the platforms of half of them do not possess at all such capabilities, while the other half has linked such features to other characteristics of their products, such as classification, life-cycle management, workflow management and extraction of dates from the content of documents (although it is to be noted that linking appraisal and retention schedules implementation with classification and life-cycle management is an approach fully acceptable in archives and records management) or, in one case, to the deployment of additional modules to be added to the original platform.

Archival remarks

In terms of records aggregation or re-aggregation, the promises of AI are not very encouraging, as this possibility is limited to very specific cases: for document types, when the users' specifications are in place, or the structure of the content source provides basic information. Automatic or semi-automatic aggregation based on the document content is only suggested with the support of the user validation, of human-in-the-loop workflow, or when content rules are created and enabled. In most cases, even these capacities are not already developed but in the process of being developed.

Even the provenance information seems not easily recognisable by AI solutions when they have to be based on inferences and there are not very specific requirements (such as the identification of the right case-folder, the presence of a stamp, a statement clearly expressed in the record, specific metadata and/or classification elements). Also the reconstitution of the archival bond – when lost or not explicitly defined – is not a simple and easy activity to be dealt with by AI solutions without the significant help of users and/or consistent descriptive information available, and, in any case, it implies more investments, not yet supported by the market.

A similar observation can be made for appraisal, not really developed by the companies that accepted to be interviewed. The only positive answer related to a product that can be used for appraising records or implementing retention schedules admits that the process requires the input and feedback of humans.

In conclusion, we have noticed a general cautious approach in all the replies when the questions were related to contextual relations. Of course, these remarks imply further analysis, as many other market proposals for archival and records management are not in line with this evaluation. The reasons for this gap could depend on the strict parameters we have adopted for selecting the market solutions. In any case, it attests to the complexity of the archival functions, which at the moment cannot be easily reduced and removed by an automatic approach, but only supported by the AI technologies through the intermediation of users and professionals. Then again, if we consider the training strategies adopted to enable and improve the capabilities of AI platforms, supervised and semi-supervised strategies posit that humans continually provide AI with labelled input and output data, which necessarily requires the intervention of experts conversant with the discipline that AI applications are intended to support. As to unsupervised strategies, first and foremost they generate outputs to be validated by humans, who are expected to give machines relevant feedback; and this of course implies the involvement of professionals knowledgeable in their area of expertise; secondly, the outcomes observed so far by research team seem to suggest that AI platforms developed by using essentially unsupervised strategies still struggle to perform complex archival tasks going beyond simple automatic indexing and categorisation.

We are not able to say, without further analysis and the conclusion of the case studies under development, which degree of professional intermediation is and will be necessary in the next and medium-term future. More effort is still required for assessing and measuring the quality and the consistency of new AI tools and their promises for automatically supporting or even substituting the human activities for classifying and aggregating records on a functional, accurate and reliable basis. This is an essential reason for our study group to focus our efforts on case studies with the aim of acquiring concrete elements for understanding how archivists could provide their support in this crucial phase of digital transformations.

Stefano Allegrezza is Associate Professor of Archival Science at the University of Bologna (Italy). He is a member of the Council of the Alma Mater Research Institute for Human-Centred Artificial Intelligence (ALMA AI) and the Director of the Summer School on "Web and social media archiving and preservation". He is the author of articles and books on the topics of records management and digital preservation.

Massimiliano Grandi is an archivist and records manager in Italy and UK for various organizations and companies. He has also taken part in InterPARES 3 (2007-2012) and contributed to the drafting of standards, terminology, and professional certification programmes. He has experience in the application of XML-based mark-up languages to the digital management and preservation of records and has also authored or co-authored articles concerning various aspects of archives and records management

Mariella Guercio is emerita Full Professor of archival science and electronic records management at the University of Urbino and Rome La Sapienza (1998-2017). She has contributed to national and international initiatives for electronic records management and digital preservation, and to EU-funded collaborative projects. She has been a member (of the ICA Programme Commission (2011-2022). She is the author of articles and books on archival science, ERMS and digital preservation.

Bruna La Sorda is an archival consultant. She was part of the working group for the translation of archival terms in InterPARES/Team Italy project. She has a leading role in the working groups for the certification of the profession as UNI standard, and for the archival services qualification. Since 2015 she is member of the national board of ANAI, the Italian national archival association, where she currently holds the position of vice-president.

Maria Mata Caravaca is the Records and Archives Manager at ICCROM (International Centre for the Study of the Preservation and Restoration of Cultural Property), where she also acts as Internal Data Processor for personal data protection. In 2017, she obtained her Ph.D. from Sapienza University of Rome with a thesis on policies and requirements for archival sedimentation. Current projects include digitization of archival resources, safeguarding of heritage samples archives, as well as participation in the InterPARES Trust AI project.

Appearance-Based Archival Science
Non-Textual Analysis and Classification of Middle-Age Parchments

by Emanuele Frontoni

Introduction

Historical parchments, acting as gateways to our past, provide invaluable insights into human communication and our rich cultural heritage. These documents, particularly those that are ancient private records, remain an epitome of human interactions in the form of agreements, indentures, treaties, and various other formal and informal transactions that have shaped societies. However, as time advances, preserving these parchments becomes increasingly challenging due to their susceptibility to physical degradation.

To tackle thfis preservation challenge, institutions like the "Archivio di Stato di Milano" (ASMi) have embarked on massive digitization undertakings. ASMi boasts an extensive collection of parchments from the 12th and 13th centuries, comprising over 130,000 documents that encapsulate nearly a millennium of history. Yet, the digitization process is not without its unique set of challenges. The ink used in these parchments, over the centuries, has deteriorated, often due to the corrosion of iron within the residual ink particles. This degradation has rendered many of these documents nearly unreadable, limiting their accessibility and usefulness.

With the advent and rapid evolution of Deep Neural Networks (DNNs), the archival world saw a shimmer of hope. Where traditional image enhancement techniques faced limitations, Deep Learning (DL) emerged as a formidable tool, capable of processing vast datasets and deciphering deteriorated writing on parchments. The evolution from conventional feature-based reading methods to DL signifies a leap in computational prowess and analytical precision.

Appearance-based signum detection

Central to the archival process at ASMi is the identification of the 'signum'. This authentication mark, used ubiquitously by notaries, is a treasure trove of information. Each distinct 'signum' provides a unique fingerprint, allowing historians and researchers to trace documents back to specific notaries, understanding their works, influence, and the breadth of their contributions. The innovative AI-driven system named PergaNet has been tailor-made to facilitate this identification. PergaNet, with its deep learning foundations, leverages appearance-based techniques to automate the extraction, recognition, and cataloging of these markings, progressively building a comprehensive database of notaries acting in Milan that spans centuries.

The intertwining of AI methodologies with archival processes represents more than just technological evolution; it signifies a paradigm shift in how we perceive, store, and analyze history. This synergistic relationship is fostering enhanced data recovery, finer historical analysis granularity, and the generation of invaluable AI datasets that echo the voices of past millennia. As we continue to push the boundaries of what AI can achieve within the archival domain, we are setting the stage for a future where our past is not just remembered but deeply understood and appreciated.

PergaNet employs DL methodologies for various tasks. The first critical task is distinguishing between the recto and verso of a parchment. This differentiation is vital since historical reconstructions are grounded in the 'signum tabellionis', a feature found only on the recto side. For this, we implement a binary classification using the VGG16 Network, renowned for its prowess in image classification.

For the following task, detecting the signa, a Convolutional Neural Network is deployed. We have chosen the YOLOv3 algorithm for its real-time processing capabilities and its computational efficiency, combined with precise detection and classification capabilities. We pretrained this network with the COCO4 dataset, a public dataset. This choice minimizes the requirement for vast amounts of training data and computational resources.

The digitized parchments (*Image 1*) are labeled using signa tabellionis, assigning an historical period to the images. The DL-based system can then detect text and signa, recognizing the notary's signum. Subsequent data analytics layers provide insights from both historical and system performance perspectives.

Image 1: Examples of Scanned Parchments dataset images. All images by the author.

The final task entails detecting and recognizing the signum tabellionis. The YOLOv3 algorithm is trained and validated using a total of 2,700 images, representing signa detection across nine notary classes. The model uses initial parameters pre-trained on the COCO dataset and is trained over 200 epochs. The model's performance is subsequently evaluated on a test set of 300 images. The manual annotations serve as the ground truth for the signa in each test image. The results (*Image 2*) suggest that our method offers substantial accuracy across different notary classifications.

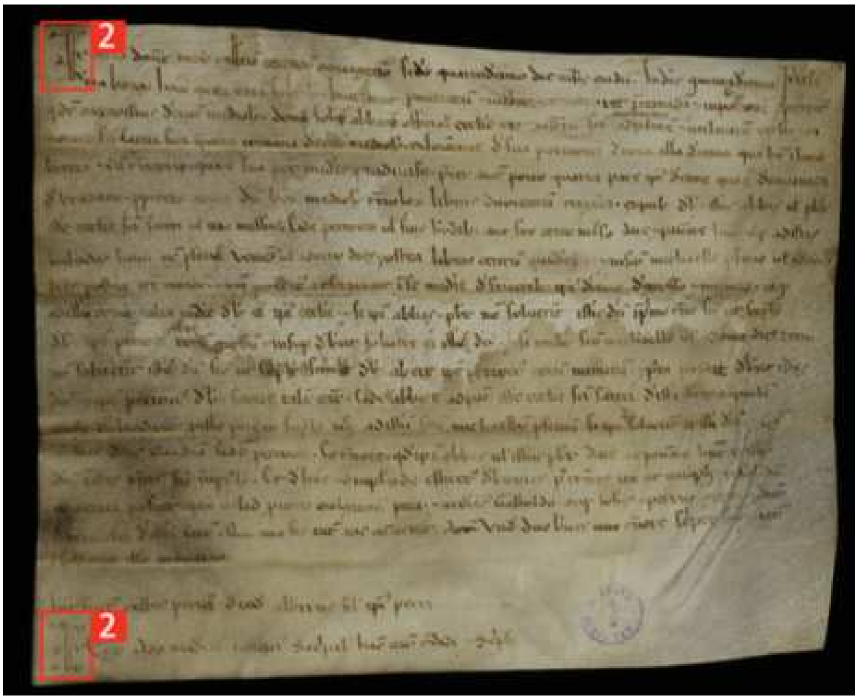


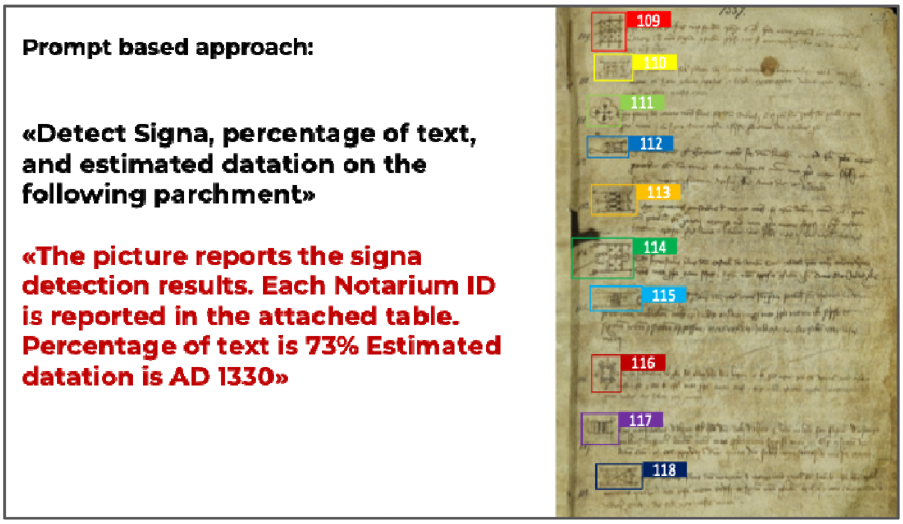
Image 2. Detection of signum tabellionis.

A prompt-based approach

Moving forward from a classical DL approach we also tried the integration of large language models (LLM) in the process of information discovery and semantic segmentation for text detection and signa classification. The prompt-based approach to image semantic segmentation leverages the capabilities of language models to interpret and generate textual prompts, which are then utilized to guide the segmentation process. This approach intertwines the advancements in natural language processing (NLP) and computer vision to enhance the signa detection process usability.

The prompt-based approach to semantic segmentation typically involves a two-step process: prompt generation and segmentation guidance. The first one uses a language model to generate descriptive prompts based on the image content or a predefined task. These prompts succinctly describe the objects, their attributes, and their relationship within the image, serving as a contextual guide for the segmentation task. The semantic segmentation model, often a DNN like U-Net or DeepLab, utilizes these generated prompts to focus on relevant features and boundaries. The prompts act as a form of conditional input, steering the model towards more accurate segmentation based on the linguistic cues.

In the example reported in *Image 3* a prompt-based approach for signa detection is reported.



Prompt based approach:

«Detect Signa, percentage of text, and estimated datation on the following parchment»

«The picture reports the signa detection results. Each Notarium ID is reported in the attached table. Percentage of text is 73% Estimated datation is AD 1330»

Image 3. Prompt-based detection of signum tabellionis and text detection.

Towards Appearance-based Archival Science

In the ever-evolving landscape of digital transformation, archives are transitioning from structured aggregations of administrative records to a data-centric model, as indicated by Moss.[25] With this shift, AI tests and challenges traditional archival principles such as respect des fonds and original order. Instead, it brings more adaptive orderings, like indexing records by content. As traditional methods discover limitations in this new domain, they can also find rejuvenation. Archivists' long-standing expertise in areas such as records provenance, transparency, and accountability becomes increasingly relevant to the AI-driven world.

The implications of AI for archiving are profound. Not only does it introduce a potential for automation in essential archival processes like metadata creation, but it also raises unaddressed concerns. For instance, AI might inadvertently introduce biases with far-reaching ethical ramifications. Discussions regarding the technical, ethical, and societal consequences of the use of AI in the archival realm are still in their infancy. The usage of AI to arrange and access archives has received substantial attention, promising an improved archival experience for diverse users. However, the impact of AI on researchers and scholars accessing these archives is still underexplored. Collaborative dialogues among archivists, AI specialists, and humanities scholars are essential to harness the full potential of AI-enhanced archives.

With digital transformation, new forms of archives, such as created in the context of social media or the Internet of Things, emerge. These new-age archives utilize AI not only for organization and access but also for their very creation. The potential of AI extends further: it can for example support democratic archival processes, highlighting records from marginalized communities, as shown by Gupta.[26] Analyzed through the

[25] Moss, Michael, David Thomas, and Tim Gollins. "The Reconfiguration of the Archive as Data to be Mined.) *ArchivariaI,* November 26, 2018, 118-151.

[26] Gupta, Abhishek, and Nikitasha Kapoor. "Comprehensiveness of Archives: A Modern AI-Enabled Approach to Build Comprehensive Shared Cultural Heritage." arXiv, August 11, 2020. http://arxiv.org/abs/2008.04541.

Records Continuum model, current work mainly emphasizes organizing and making records accessible, somewhat overlooking the continuous evolution and creation of records with AI's involvement.

Our study revealed intriguing trends. Traditional archival methods are being challenged, with AI pushing and redefining their limits. Principles like respect for original order, which once guided the organization and access to records, now represent just one method amidst the sprawling digital archives we grapple with. AI introduces versatile, dynamic structuring methods, such as organizing by content. The AI domain presents challenges that highlight the need to integrate recordkeeping expertise in the AI systems of today. These challenges, which include understanding data origin, evaluation, context, transparency, and responsibility, are all areas where archivists' expertise shines. Responding to these challenges means reimagining how we train archivists. Our data indicates a recognition of this need, but tangible solutions remain sparse. Observing AI in the archival realm, there's a surge in activities that aim to automate processes, from appraisal to metadata generation. While promising, our observation also indicates a cautionary approach towards fully embracing AI because of the mentioned biases introduced by AI and their ethical ramifications. Steps are being taken to address these concerns, but in-depth discussions on AI's technical, ethical, and societal impacts in archiving still await.

AI's potential in redefining archival access is another point of discussion. Many are rethinking the interaction with archives and their information, thanks to a wealth of contributions in the field. While fully integrating AI in archival systems remains a goal, our findings show promising examples of AI serving diverse user needs. A future exploration area is understanding AI's impact on researchers accessing these archives. There's potential for AI-driven tools that can elevate scholars' interaction with archives. This evolution necessitates collaboration between archivists, AI experts, and humanities scholars. Digital evolution is expanding what we consider a record. Emerging digital archives are AI-reliant for organization, access, besides their very inception. AI also promotes a more inclusive archival approach, bringing forward records from marginalized groups. These new-age archives demand innovative thinking and the ability to manage vast data volumes.

Emanuele Frontoni is Full Professor of computer science with the University of Macerata (Italy) and Co-Director of the VRAI Vision Robotics & Artificial Intelligence Lab. His research interests include computer vision and artificial intelligence with applications in robotics, video analysis, human behaviour analysis, extended reality and digital humanities. He is a member of the European Association for Artificial Intelligence, the European AI Alliance, and the International Association for Pattern Recognition.

The Crucial Role of Paradata in AI Governance

by Patricia C. Franks

The introduction of new technology carries with it elements of risk. Artificial Intelligence (AI) can pose risk to humans - such as a loss of human connections and potential job loss. It can pose risks to the company employing it - including a lack of protection of data privacy and reputational damage. In addition, it can pose risk to society as a whole - as in the case of social surveillance and autonomous weapons.

AI developers face lawsuits based on the data used to train their models. OpenAI, developer of ChatGPT, and its partner Microsoft are being sued by both the *New York Times* and a group of well-known authors for using their intellectual property as training data.[27] And Claude AI developer Anthropic is being sued by Universal Music Group (UMB) and other publishers for "allegedly" feeding the songs of artists they represent into its AI models without permission.[28]

Risks are also incurred by businesses employing AI. Lawsuits over unacceptable outcomes of AI tools have been filed. United Healthcare, for example, is being sued over an algorithm that overrode physician recommendations and resulted in wrongfully denying coverage for stays in extended care facilities.[29] Sunglass Hut, Macy's is being sued by a grandfather for damages after facial recognition led to his wrongful arrest, imprisonment, and sexual assault in jail.[30]

The unfavourable outcome of the use of an AI model can damage the reputation of an organization even if it is not the subject of a lawsuit. Recently, Google's AI chatbot, Gemini, was in the news for unrealistic images portraying historical characters and for controversial responses to questions. Google lost more than $90 billion in market value. While the value of the stock has recovered, as Melius Research analysts Ben Reitzes and Nick Monroe wrote in a note to clients:

> "The issue for the stock is not the debate [over Gemini] itself, it is the perception of truth behind the brand. Regardless of your view, if Google is seen as an unreliable source for AI to a portion of the population, that isn't good for business."[31]

[27] Grynbaum, M., and Mac, R. (2023, December 27). The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work, *The New York Times*, https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

[28] Nelson, J. (2023, October 18). Universal Music Group Sues Anthropic Claiming 'Widespread Infringement,' *Emerge*, https://decrypt.co/202314/universal-music-group-umg-anthropic-ai-copyright-lawsuit

[29] Claburn, T. (2023, November 15). UnitedHealthcare's broken AI denied seniors' medical claims, lawsuit alleges," https://www.theregister.com/2023/11/15/unitedhealthcare_ai_medicine/#

[30] Business Wire. Man Sues Sunglass Hut, Macy's over False Imprisonment, Sexual Assault, https://www.businesswire.com/news/home/20240122482901/en/Man-Sues-Sunglass-Hut-Macy%E2%80%99s-over-False-Imprisonment-Sexual-Assault

[31] Saul, D. (2024, February 26). Google's Gemini Headaches Spur $90 Billion Selloff, *Forbes*, https://www.forbes.com/sites/dereksaul/2024/02/26/googles-gemini-headaches-spur-90-billion-selloff/?sh=6c7d705372e4

AI Governance

Artificial Intelligence Governance is the most effective method to mitigate risks associated with the design, development, and deployment of AI, and to instil trust in the end user.

AI Governance "refers to the guardrails that ensure AI tools and systems are and remain safe and ethical. It establishes the frameworks, rules and standards that direct AI research, development and application to ensure safety, fairness and respect for human rights."[32]

AI Governance resources can be divided into 5 categories, as shown in Figure 1: principles, frameworks, laws and policies, voluntary guidelines, and standards and certifications. All are essential to developing an AI Governance strategy.

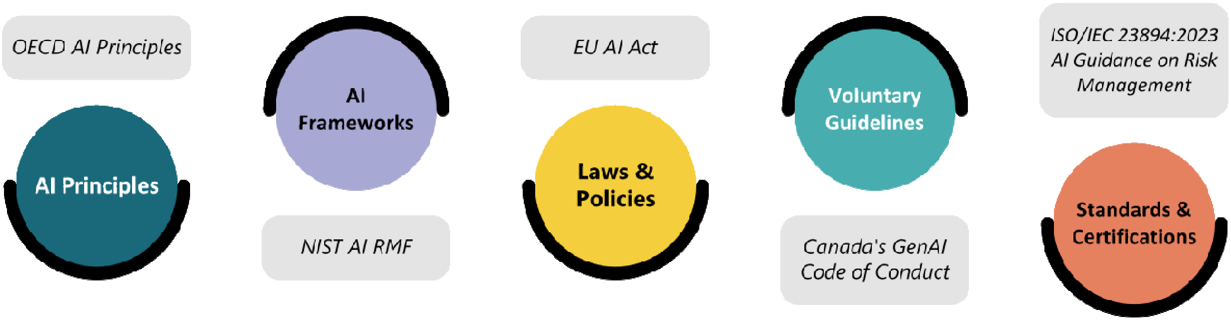

OECD AI Principles
AI Principles
AI Frameworks
NIST AI RMF
EU AI Act
Laws & Policies
Voluntary Guidelines
Canada's GenAI Code of Conduct
ISO/IEC 23894:2023 AI Guidance on Risk Management
Standards & Certifications

Figure 1. Categories of AI Governance Resources and an example of each. Image by author.

AI Principles

The *OECD AI Principles* promote AI that is innovative and trustworthy and that respects human rights and democratic values. They focus on how governments and other actors can shape a human-centric approach to trustworthy AI. Adopted in May 2019, the *AI Principles* set standards that are practical and flexible. They represent a common aspiration for adherents to the OECD Principles: 38 member countries and 8 non-members.[33]

AI Frameworks

Several frameworks exist to provide guidance for managing AI risks. The *AI Risk Management Framework* (AI RMF 1.0) published in January 2023 by the U.S. National Institute of Standards and Technology (NIST) refers to systems that generate objectives, recommendations, or decisions that influence real or virtual environments.[34]

[32] Mucci, T., and Stryker, C. (2023, November 28). What is AI governance? *IBM*, https://www.ibm.com/topics/ai-governance#

[33] Russell, S., Perset, K., and Grobelnik, M. (2023, November 29). Updates to the OECD's definition of an AI system explained. OECD.AI, https://oecd.ai/en/wonk/ai-system-definition-update

[34] NIST. AI Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework

Four tasks recommended to manage risk are govern, map, measure, and manage. Examples of the actions and decisions that should be documented to meet the guidelines are:

- *Govern*: Legal and regulatory requirements involving AI are understood, managed, and documented.
- *Map*: the AI system's features and capabilities that require human oversight must be documented, and training materials must be developed.
- *Measure*: Documentation of the AI system's data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints, and metadata must be made.
- *Manage*: Responses to the AI risks deemed high priority as identified by the Map function must be developed, planned, and documented.

A companion *AI RMF Playbook*, which will be updated semiannually, suggests actions, references, and related guidance to achieve the desired outcomes. A supplement to the *AI RMF Framework* is slated for release in June 2024 to provide guidance for Generative AI.

Laws and policies

The best example of AI laws is the *EU AI Act* approved by parliament on March 13, 2024, to ensure safety and compliance with fundamental rights, while boosting innovation.[35] The EU Act takes a risk-based approach to AI technologies, categorizing risks into four levels.

Unacceptable-risk systems that threaten fundamental rights, such as biometric categorization systems that use sensitive characteristics (e.g., political, religious, philosophical beliefs, sexual orientation, race), and untargeted scraping of facial images from the Internet or CCTV footage to create facial recognition databases. These systems are banned. Exemptions are possible for use by law enforcement with judicial authorization.

High-risk systems pose "significant potential harm to health, safety, fundamental rights, environment, democracy, and the rule of law."[36] Examples include certain critical infrastructures (e.g., energy and transport); essential private and public services (e.g., healthcare and banking); and systems influencing democratic processes (e.g., fair elections). Compliance obligations such as maintenance of use logs and human oversight are mandated.

Limited-risk systems include chatbots and certain emotion recognition and biometric categorizations systems as well as those that generate deep fakes. Minimal transparency

[35] Mackrael, K., and Schechner, S. (2024, March 13). European Lawmakers Pass AI Act, World's First Comprehensive AI Law, The Wall Street Journal, https://www.wsj.com/tech/ai/ai-act-passes-european-union-law-regulation-e04ec251

[36] European Parliament (2023, September 12). Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI, https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

obligations include informing users they are interacting with an AI system and marketing synthetic audio, video, text, and image content as artificially generated or manipulated.

Minimal or no risk systems can be used freely; however, voluntary codes of conduct are encouraged. Examples of these systems are AI-enabled recommendation systems, spam filters, and video games.

Laws and regulations are external, but employees must comply with internal requirements as well. AI policies should be developed and shared that cover data privacy, bias, transparency, and accountability and provide guidance on how to handle ethical dilemmas.

Voluntary Guidelines

In September 2023, Canada launched a voluntary code of conduct for companies developing generative AI. The code includes measures for accountability, safety, fairness and equity, transparency, human oversight and monitoring, and validity and robustness.[37]

AI Standards

The International Organization for Standardization (ISO) has published numerous AI-related standards. Two standards related to the topic of AI Governance are ISO/IEC 23894:2023 - Guidance on Risk Management, and ISO/IEC 38507:2022 - Governance Implications for organizations. Additional standards address topics of AI such bias, neural networks, and use cases.[38]

AI Certifications

AI certifications on the horizon! In January of 2024, an initiative was launched to evaluate and certify AI products as copyright compliant. The *Fairly Trained* label will be given to AI companies that have obtained consent for the data they use to train AI systems.[39]

AI Lifecycle

The AI Lifecycle as visualized by the General Service Administration of the U.S. Government is comprised of three phases:

- Design: understand the problem, gather data, prepare data for model development.
- Develop: Train and test the mode using data gathered.
- Deploy: Move the model to production, implement it, gather feedback on the output.[40]

[37] Government of Canada (2023, September). Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems

[38] International Organization for Standardization (ISO). Artificial Intelligence, https://www.iso.org/sectors/it-technologies/ai

[39] Huet, E. (2024, January 22). AI certification program verifies systems are 'Fairly Trained,' *The Seattle Times*, https://www.seattletimes.com/business/ai-certification-program-verifies-systems-are-fairly-trained/

Each stage of the AI life cycle involves actions taken and decisions made by AI or in conjunction with human agents that must be documented as evidence of responsible and accountable use of AI.

Paradata

Perhaps the principal role of archivists and records managers in AI governance will be as the stewards of the documents, records, and data that must be retained as evidence of responsible and accountable AI. As posed by Jenny Bunn, "If business is no longer to be transacted only by human beings, but also by AI agents, or some combination of the two, what will evidence of those transactions look like, what will the record be?"[41] And as Norman Mooradian suggests, "Defining an AI record and developing methods for capturing AI records is a project the profession should take on."[42]

Consequently, a team was convened in fall of 2021 to address this challenge. The primary research question identified was:

> If an AI technique is used to facilitate or automate an archival, recordkeeping, or other process, how much of that AI technique, its code, the data (probably a subset of existing records) we use to train it, test cases and test results to examine its efficacy, its parameters and their values at or over the time of application, the technical environment in which it is executive, and the records it (the AI technique) is applied to for automation purposes, should be preserved?

To answer this question, an investigation into the type of auxiliary information recorded in various fields - such as statistical research, social science, and visual heritage - was conducted. The term paradata emerged across these disparate fields to refer to processual data within their discipline-specific contexts. It was determined that paradata could be considered an all-inclusive umbrella to aggregate the various pieces needed to record evidence of AI processes. Therefore, the following definition of paradata for the AI process was adopted:

> Paradata is ***the information about the procedure(s) and tools*** used to create and process information resources, along with ***information about the persons*** carrying out those procedures.

[40] The AI Lifecycle. *Source*: GSA AI Guide for Government, https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/

[41] Bunn, J. (2020), "Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI)", *Records Management Journal*, Vol. 30, No. 2, pp. 143-153. https://doi-org.libaccess.sjlibrary.org/10.1108/RMJ-08-2019-0038

[42] Mooradian, N. (2019). AI, Records, and Accountability, ARMA *Magazine*, https://magazine.arma.org/wp-content/uploads/2019/12/ARMA-Magazine-AIEF-Special-Edition.pdf

Paradata can be thought of as processual documentation. It must document the full scope of application and context of use—not just the algorithm itself. Explainable AI (XAI) clarifies why a given tool produced a given output from a given set of inputs. But paradata is necessary to explain why, how, and to what effect a given tool was used in a particular context.

There is also a difference between metadata and paradata, although the two may at times overlap. Metadata can be thought of as data about the information resource for the purposes of documenting, describing, and preserving or managing that resource. Paradata is information about the AI process that enables processual insight, transparency, and accountability. Examples of documentation categorized as technical or organizational paradata are shown in *Table 1*.

Table 1. Examples of paradata.

Technical Paradata	**Organizational Paradata**
AI Model (tested and selected)	AI policy
Evaluation & performance metrics	Design plans
Logs generated	Employee training
Model training dataset	Ethical considerations
Training parameters for the model	Impact assessments
Vendor documentation	Implementation process
Versioning information	Regulatory requirements

The following message was shared with participants at a 2023 Research Forum held by the Society of American Archivists:

> "As the use of artificial intelligence continues to grow and algorithms and models become more complex, so do the challenges of explaining, justifying, and providing evidence of the actions and decisions carried out with little or no human intervention. It is, therefore, incumbent upon archivists and records managers, long accustomed to documenting actions and decisions of humans, to lend their expertise to documenting the actions and decisions of AI systems."[43]

Technical paradata documenting the actions taken and decisions made throughout the AI process is naturally collected by developers or the systems themselves to improve system design. AI governance requires the creation and/or capture of organizational paradata as well. Both forms of Paradata should be captured and preserved so that it can

[43] Franks, P.C. (2024, March). In the Pursuit of Archival Accountability: Positioning Paradata as AI Processual Documentation," *Society of American Archivists - 2023 Research Forum*, https://www2.archivists.org/sites/all/files/Franks_In%20the%20Pursuit%20of%20Archival%20Accountability.pdf

be used to explain, justify, and provide evidence of the actions and decisions carried out by AI and humans throughout the AI Process.

Conclusion

Every organization should begin to discuss the topic of AI Governance. A risk-based approach is recommended, and high-impact, high-risk AI implementations should be prioritized. Guidance in the form of laws, regulations, frameworks, and standards must be monitored and applied to the AI process as necessary.

While guidance documents refrain from the use of the term "record" in most cases, the terms document and documentation are widely used and imply that records of the AI process be captured and preserved. Some of the documentation will be automatic as part of the AI system; some will be human-created prior to or after the creation and implementation of the AI system. According to Cameron and Hamidzadeh,[44] "the information community must develop the concepts and vocabulary to describe its needs rooted in the field's professional values." Paradata is a term newly added to describe information about the AI process. The archival perspective supports the capture and preservation of paradata—and is necessary to ensure the AI process is documented in a way that preserves the characteristics of authoritative records: authenticity (i.e. identity and integrity), reliability, accuracy, and usability.

Dr. Patricia C. Franks teaches courses in Enterprise Content Management and Digital Preservation in the School of Information at San José State University, California, USA. She is a Certified Archivist, Certified Records Manager, and Information Governance Professional. She authored the book Records and Information Management *and edited* The Handbook of Archival Practice. *Her research interests lie in emerging technologies and records and information management.*

[44] Cameron, Scott and Hamidzadeh, Babak, Preserving Paradata for Accountability of Semi-Autonomous Ai Agents in Dynamic Environments: An Archival Perspective. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4681230

Hybrid Agency and Real-Time Systems: Paradata for Accountability of AI Systems within and beyond Traditional Archives

by Scott Cameron[45]

The need for paradata in the archival profession is rooted in the contention that archivists should record how they process the records in their fonds, and how this processing may influence the ways in which users experience and encounter them. In this sense, while paradata is a new term, it is based in enduring principles of transparency and impartiality within the archival profession. As new technologies emerge, new vocabulary may be occasionally necessary to describe the challenges that they present. In the case of paradata, the term merely describes the archivist's processual documentation, the information necessary to communicate the influence that the archivist's actions and decisions, mediated through the tools they employ, may have had on the record(s) which researchers encounter. Paradata thus documents archival processes that may otherwise be opaque to external observers, and avoids or at least limits the risk of the black box problem. The black box problem has been much discussed in the context of artificial intelligence, machine learning and computing. Should archivists introduce these tools into their operational routines, it behoves them to avoid introducing also the pitfalls of complex computing tools into trusted repositories. However, even prior to the introduction of machine learning, the work of the archivist has hardly been universally transparent to external users. The use of paradata is thus an opportunity for archivists not just to make transparent, accountable, and fruitful applications of AI tools in archives, but also to use computerized tools to make accountable and transparent much of their work, developing better and more effective means of documenting archival decisions made throughout the life of the materials in their care in a systemic manner at scale. Paradata thus provides an opportunity not just to document computer processes, but also the actions of the people and organizations implementing them.

Since the use of advanced computational tools risks the introduction of complex or opaque processes into accountable contexts entailing significant risk, provision of clear documentation of actions and of the agents responsible for them is a precondition for accountable system operations. However, this elevated documentation needs reveal the shortcomings of automated systems mostly at the point where they intersect with human systems. We suggest that the activities requiring the greatest burden of documentation will frequently be in systems and contexts that combine or blend human and AI agencies. Even when an AI tool is a complete and perfect black box, offering no insight into its decision processes, as long as the limits of the tool's operating environment are known, then tracing responsibility for its actions is relatively uncomplicated. Whether the AI tool

[45] *Acknowledgements*: This piece draws upon a forthcoming paper by Scott Cameron and Babak Hamidzadeh. The author would like to acknowledge Dr. Hamidzadeh's contributions to this work, the support of our colleague Dr. Patricia Franks in the paradata working group, as well as the support of Drs. Luciana Duranti and Corinne Rogers in the InterPARES Trust AI project.

operates with or without explanation, no question exists as to the tool's actions within its operating context. The same is true of human agents; as long as the individuals operate within a prescribed field of action, whether they provide explanation for their actions or not requires relatively little documentation, provided that the individuals themselves can provide an account for their decisions and ultimately be held responsible for them.

In contrast, in scenarios where AI tools share responsibility with human agents, defining the boundaries of responsibility becomes the primary functional requirement of effective records systems. Increasingly common are AI systems that operate in dynamic, real-time environments and divide decision-making capacities between autonomous system functions and human responsibilities. A hybrid agency problem emerges as decision-making capacities are amalgamized between the two parties. A critical voice may note that, within most juridical contexts, responsibility for an AI system may be clearly determined either by law or by contract between responsible parties. However, while clear assignments of responsibility may reduce the burden of documentation in some cases, other practical applications, such as quality control, may demand extensive and granular records availability, as those designing, marketing or implementing AI systems may need to prove their due diligence.

To illustrate the contexts in which operational paradata may prove a key part of the records available, two examples are presented to show the increased burden of documentation necessary in shared agency environments: the self-driving car and the digital twins used in built infrastructure management. As autonomous vehicles (AVs) have become increasingly viable, the limitations of nominally autonomous systems have come into clearer view. Rather than dividing vehicles into autonomous and non-autonomous categories, the Society of American Engineers (SAE) uses a numerical scale ranging from 0 to 5 based on whether a vehicle is entirely autonomous or blends autonomous systems with conventional human control. Many new vehicles on the road already incorporate at least some autonomous features and fall between levels 0 and 2 on the SAE scale: blind spot warnings, adaptive cruise control or lane centring features all introduce limited autonomous features as the driver maintains higher-level control of the system. At level 3, vehicles may operate autonomously but require drivers to take control at a moment's notice in emergencies. At level 4 vehicles are entirely autonomous within circumscribed zones, and at level 5 vehicles are autonomous with no geographic limitations (SAE, 2021). At this time, level 4 vehicles are used in extremely limited spaces, and level 5 vehicles exist only on paper (Kosuru and Venkitaraman, 2023). As truly autonomous vehicles have remained elusive, the increasing and persistent prevalence of semi-autonomous systems illustrates the ongoing hybridity of nominally autonomous systems, even as AI capacities increase in this sphere.

The second example describes digital twin systems, a form of predictive control systems based in extensive spatial data infrastructure that illustrates another example of the hybridity of existing AI systems. Digital twins are complex computerized control systems for real-world infrastructure, defined formally as "an ecosystem of multidimensional and interoperable subsystems made up of physical things in the real-world, digital versions of those real things, synchronized data connections between them and the people, organizations and institutions involved in creating, managing, and using these" (Frontoni et al., 2022, p. 6). Vancouver's YVR International Airport provides an illustration of a digital twin system employed in an accountable context. YVR's digital twin presents a complex, hybrid agency environment. A current application of the system is to manage flows of pedestrian traffic within the airport. Modelling and predicting crowd behaviour using the model allows YVR to direct pedestrians efficiently and minimize delays in passenger throughput within the airport (YVR, 2023). Changes such as those to electronic directional signage may be implemented directly by the digital twin system, whereas other changes, such as moving cordons or barriers, may need to be reviewed and enacted by humans. As the system's capacities improve, other potential uses may include the identification of security issues, real-time monitoring of maintenance work and facility conditions, direction of air or ground traffic at the airport, and the simulation of emergency response plans. All of these scenarios will involve executive decision-making processes merging the tool's decision processes with those of the facility staff, illustrating another case of the hybrid integration of automated systems within existing traditional infrastructure. Rather than imagining AI as fully replacing human capacities, we are more likely to see AI processes integrated into traditional ones, introducing new hybrid agencies and the risk of blurred responsibilities. These systems will pose challenges for records creation and preservation, as large, complex and opaque datasets documenting dynamic processes will be introduced alongside more traditional forms of documentation.

What records then might be necessary to document real-time hybrid agency systems, and what might these records look like? What information is generated throughout the AI process that may offer significant insights into its outputs or operations? What is clear is that high frequency real-time data recording the behaviour of complex systems in accountable environments will become necessary as the scale of AI tools' implementation increases. In cases where archivists have the opportunity not just to maintain records after their life cycle has ended but to offer input into records creation before systems are implemented, the obfuscations introduced by AI may be mitigated by close attention to the possible types of paradata which may be gathered and preserved. As a neologism, empirical researchers coined the term "paradata" to describe information generated during an information artefact's creation that offers insight into the agents, processes and decisions involved in the process. While paradata may include information recorded incidentally or later deduced inferentially throughout a related information resource's creation process, systematic approaches to recording, preserving, and disseminating paradata in empirical research fields are increasingly common. Understanding these paradata in relation to the primary information object provides insight into the processes which led to the creation of that information resource. Rather than presenting the primary information object as static, paradata allow the primary information resource to emerge as the product of dynamic and contingent processes.

Assessing the paradata which might be gathered and preserved to elucidate and record for posterity the actions of dynamic hybrid systems is a complex problem. We suggest that close analysis of each system in question is necessary to understand the records needs. Cases where AI systems offer executive or advisory roles to augment or partially replace human capacities present significant records needs when implemented in accountable contexts. The operation of complex AI systems in dynamic environments is contingent on sensors, a control system, and actuators to create change in the real world; in each case, these subsystems combine to produce real changes in the world. Breaking down them into subcomponents underlying their hybrid decision processes can identify the subsystems that may generate paradata necessary for preservation. Real time systems generate and act upon swathes of data that are not always preserved. For the digital twin or autonomous vehicle systems described, the table below outlines relevant paradata produced by the system's sensors, controllers and actuators, and the evidence which may be generated of the system's effectuated changes upon the real world. In either system, this information may prove to be necessary for preservation to account for the systems' operation.

System subcomponents and necessary system paradata for preservation	
What is documented	Example paradata identified as preservation targets
Sensor input	Log of sensor data (speedometer, counters, GPS data, steering mechanisms, etc.); Camera footage used for computer vision systems
Controller	Log of control directions from both human and system agents; Relevant settings of control system; Intermediary subprocess data leading up to a decision; Post-facto AI explanations of these processes; Log of warning notifications and control handover notifications
Actuators	Log of physical system's actions; Log of messages communicated from system to human controller and external parties
Effects	Real world system actions recorded through log of sensor data; camera footage; third-party evidence

While vast quantities of data may be produced by dynamic, real-time systems, archival precedents may provide useful models for approaching the tasks of selecting the data necessary for long-term preservation as electronic records. For instance, in 2006, InterPARES 2 studied the records needs of the City of Vancouver's dynamic geographic VANMap system. As the system tracks built city infrastructure over time and is used for municipal planning decisions, InterPARES recommended that preservation of the data held within the system which formed the basis of staff decisions was necessary for accountability (Duranti and Thibodeau, 2006, pp. 43-44, 65-66). Although AI hybrid agency systems may present an increased system dynamicity and the complications of hybrid agency relatively to VANMap, the prescription that the moment and basis of a decision is the key preservation target in a dynamic system remains valid. The entire dataset generated throughout the operational lifetime of a dynamic AI system will rarely be worthy or feasible for preservation by itself; rather, a record of the system's decisions and actions accompanied by the paradata underlying those processes will in most cases be a sufficient and illustrative record. As to the preservation of these paradata, their relationship with the primary record is essential to its meaning. Without preserving the relationship between the evidence of a decision and the evidence of the processes leading to that decision, the evidence produced by dynamic AI systems will return to the territory of the black box, and will remain unknown to those who attempt to track the actions of the hybrid agencies responsible for their implementation and use.

With increasing frequency, real-time AI systems have infiltrated high-risk accountable environments with inherent records needs often increased by the obscurity of AI systems. With a close analysis of the nature of these systems and an understanding of the risks of the black box and hybrid agency problems, the systems at hand do not present fatal challenges to accountability. Archivists need to understand the nature and increasing complexity of AI systems, their prevalence in accountable high-risk environments, and the growing challenges that the volume and interconnectedness of the data produced by these systems entail. As these systems pose inherent challenges to those intending to document their operation using electronic records, archivists are well-situated to offer their expertise. Understanding the risks which transient data pose and the opportunities that identifying and preserving relevant paradata provide will well-equip those intending to responsibly apply AI tools moving into the future.

References

Duranti, L., and Thibodeau, K. (2006). The Concept of Record in Interactive, Experiential and Dynamic Environments: The View of InterPARES. In *Archival Science* 6 (1): 13-68. https://doi.org/10.1007/s10502-006-9021-7

Frontoni, E., Paolanti, M., Lauriault, T.P., Stiber, M., Duranti, L., and Abdul-Mageed, M. (2022). Trusted Data Forever: Is AI the Answer? *ArXiv* preprint. https://doi.org/10.48550/arXiv.2203.03712

Kosuru, V.*S.R.,* and Venkitaraman*, A.K. (2023).* Advancements and challenges in achieving fully autonomous self-driving vehicles. In *World Journal of Advanced Research and Reviews* 18(01). https://doi.org/10.30574/wjarr.2023.18.1.0568

SAE (2021). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104.*

YVR (Vancouver Airport Authority) (2023). YVR and Unity accelerate digital transformation in aviation with YVR's Digital Twin platform. http://www.yvr.ca/en/media/news-releases/2023/yvr-and-unity-accelerate-digital-transformation-in-aviation

Scott Cameron holds a joint MAS and MLIS from the University of British Columbia and an MA in History from the University of Toronto. He has contributed to the InterPARES project's paradata working group since 2022 and published on topics of records and accountable artificial intelligence. He is currently employed as a Metadata and Standards Analyst at the Bank of Canada. The views expressed in this paper do not reflect those of his employer.

Modelling AI-Assisted Digitization of Documentary Heritage Materials

by Eng Sengsavang and Željko Trbušić
Contributors: Shadreck Bayane, Marina de Souza, Kailey Fukushima, David Iglesias, Tomislav Ivanjko, Adam Jansen, Petra Lovric, Marta Riess, Hrvoje Stancic, Goran Zlodi

Introduction

In a variety of spaces within organisations and archives, in museums and libraries, even in private dwellings, rows and piles of archival materials sit soundlessly, sometimes carefully or abruptly plucked from their resting places, brought into the light for consultation by researchers or those who care for archives. Since the global COVID-19 pandemic, the demand for digital resources has increased dramatically. The conversion of physical artefacts into digital form has, in parallel, increased to keep pace with the flow of demand, but not only. The creation of digital copies of ageing, deteriorating, or endangered archives has also become an important conservation strategy, a type of insurance against the exigencies of time - as much, perhaps, as it is a balm for the anxiety of archivists in charge of their care, or a promise made in response to the ardour of researchers invested in their long-term preservation.

While a digital copy can never replace the original, and issues around how digital copies are made public and accessed are rich topics for discussion and debate, there is no question that digitization - the process of creating a digital representation of a physical object - creates backup copies of vulnerable materials, and reduces wear and tear on the original artefacts. It is also complex and resource-intensive, involving multiple processes, activities, expertise, and technologies. The increasing popularisation of artificial intelligence has added yet another layer of complexity to a process that may appear simple, yet in reality is anything but. Our research, entitled "AI-Assisted Digitization of Archives and Documentary Heritage Materials,"[46] draws from a variety of digitization standards and best practices to document the phases and activities of archivally-oriented digitization practices - that is, digitization that aims to create "faithful reproductions" (FADGI, 2023, p. iv) of physical records to help preserve, over the long-term, both the original record and its digital surrogate. The model outlines the various points in the digitization process that may be supported by artificial intelligence. On the basis of the identified phases and activities and points of AI intervention, we created an interactive visual model, simply called the "sunburst model." To supplement our understanding of how organisations are using AI in their digitization processes, we also sent out a survey, "Digitization and Artificial Intelligence for Archives and Documentary Heritage

[46] See "AI-Assisted Digitization of Archives and Documentary Heritage Materials (RA03)" at https://interparestrustai.org/trust/about_research/studies

Materials," aimed at measuring the impact of AI on organisations undertaking digitization of their records.[47]

Modelling AI-assisted digitization

To show at what points, and how, artificial intelligence may support digitization processes, it was necessary to first identify, in a comprehensive way, the main phases and activities undertaken when digitising archives. The motivation to create a visual model was born from a need to "see" and capture the entire digitization process from a birds-eye view. The model was envisioned as a tool for archivists and digitization specialists who are considering or may wish to integrate AI capabilities into their digitization workflows.

The model was created by consulting various digitization standards and guidelines, among others, FADGI's *Technical Guidelines for Digitizing Cultural Heritage Materials* (2023) and ISO's *Information and documentation - Implementation guidelines for digitization of records* (2010), to ensure we captured as comprehensively as possible the activities involved in digitization. At the same time, the study team considered whether AI could intervene in each activity. This bottom-up approach was both organic and practical, and addressed one of several study goals, namely to create a non-prescriptive visual abstraction of the digitization process across three generalised stages - pre-digitization, digitization, and post-digitization - to serve as a starting point from which to explore how AI may be or is currently being used in digitization processes. Another main study goal is to analyse the potential benefits, risks, and biases of using AI during digitization of documentary heritage materials, while being informed by ethical and responsible approaches to the use of AI, which we will explore in further phases of the study.

The model was split into hierarchical segments of activities and corresponding sub-activities, where applicable. Each segment was marked as belonging to one or more of the three general digitization stages (pre-digitization, digitization, or post-digitization). The activities and sub-activities within each stage are non-prescriptive and are not represented sequentially, as they may be iterative, and in consideration of the fact that no single digitization project proceeds in exactly the same way. For example, metadata management was identified as a main activity potentially taking place across all three digitization stages, and involving the following sub-activities: metadata gathering, metadata creation, metadata enrichment, cataloguing and description, and metadata validation/quality control. We determined that AI may potentially intervene in all metadata sub-activities. Moreover, we identified the types of AI models or tools that could be relevant in each sub-activity, such as OCR/HTR (optical character recognition/handwritten text recognition), computer vision, and audio speech recognition for metadata creation, enrichment, cataloguing and description and quality control. In this way, the potential applicability of AI was considered for each activity and sub-activity. Finally, visual

[47] The survey report is forthcoming and will be published on the InterPARES AI website: https://interparestrustai.org/

representations were created using two modelling formats - across functional chart (*Figure 1*) and a sunburst diagram (*Figures 2* and *3*).

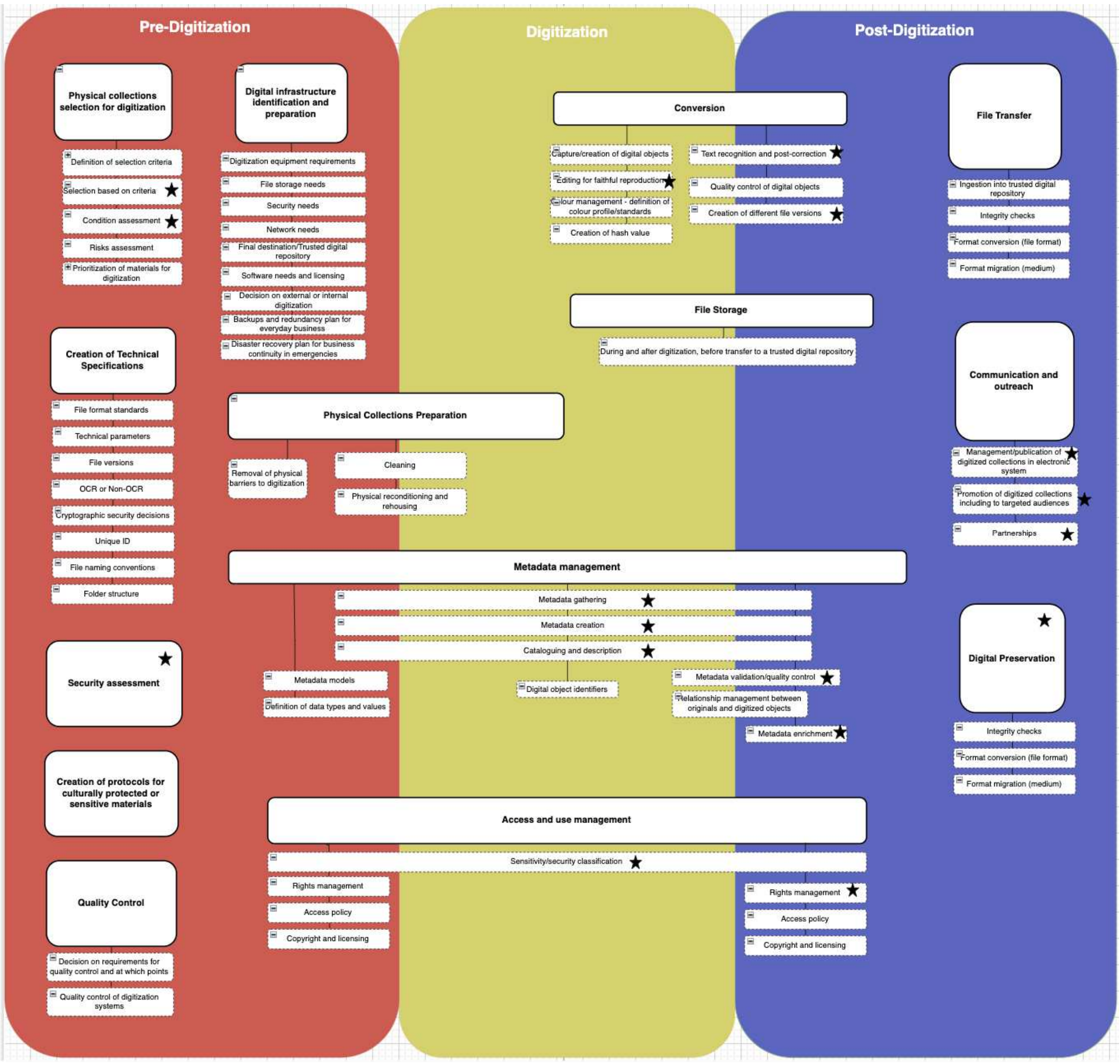
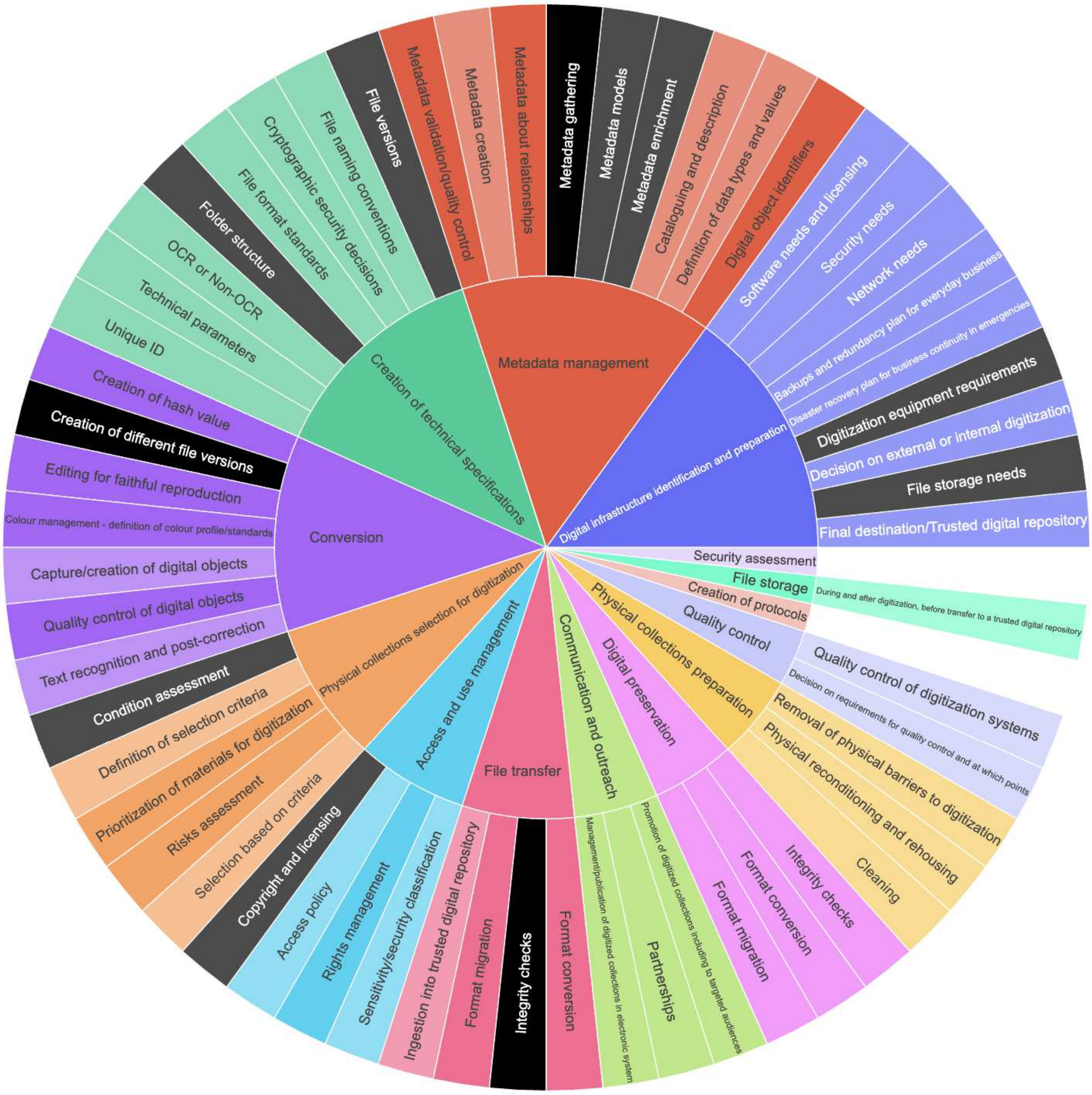

Pre-Digitization
Digitization
Post-Digitization
Physical collections selection for digitization
Definition of selection criteria
Selection based on criteria ★
Condition assessment ★
Risks assessment
Prioritization of materials for digitization
Digital infrastructure identification and preparation
Digitization equipment requirements
File storage needs
Security needs
Network needs
Final destination/Trusted digital repository
Software needs and licensing
Decision on external or internal digitization
Backups and redundancy plan for everyday business
Disaster recovery plan for business continuity in emergencies
Creation of Technical Specifications
File format standards
Technical parameters
File versions
OCR or Non-OCR
Cryptographic security decisions
Unique ID
File naming conventions
Folder structure
Security assessment ★
Creation of protocols for culturally protected or sensitive materials
Quality Control
Decision on requirements for quality control and at which points
Quality control of digitization systems
Conversion
Capture/creation of digital objects
Editing for faithful reproduction ★
Colour management - definition of colour profile/standards
Creation of hash value
Text recognition and post-correction ★
Quality control of digital objects
Creation of different file versions ★
File Storage
During and after digitization, before transfer to a trusted digital repository
Physical Collections Preparation
Removal of physical barriers to digitization
Cleaning
Physical reconditioning and rehousing
Metadata management
Metadata gathering ★
Metadata creation ★
Cataloguing and description ★
Metadata models
Definition of data types and values
Digital object identifiers
Metadata validation/quality control ★
Relationship management between originals and digitized objects
Metadata enrichment ★
Access and use management
Sensitivity/security classification ★
Rights management
Access policy
Copyright and licensing
Rights management ★
Access policy
Copyright and licensing
File Transfer
Ingestion into trusted digital repository
Integrity checks
Format conversion (file format)
Format migration (medium)
Communication and outreach
Management/publication of digitized collections in electronic system ★
Promotion of digitized collections including to targeted audiences ★
Partnerships ★
Digital Preservation ★
Integrity checks
Format conversion (file format)
Format migration (medium)

Figure 1. Cross-functional chart of digitization phases and activities. All figures by "AI-Assisted Digitization of Archives and Documentary Heritage Materials (RA03)", InterPARES Trust AI.

Interactive sunburst model

A "sunburst" representation of the model involved building an animated diagram using Plotly, an interactive, open-source, browser-based graphing library. The diagram consists of three layers. The first layer (*Figure 2*) represents the whole model, with all activities and sub-activities, and enables users to click on a desired activity to explore additional information found in the second (*Figure 3*) and third layers (not shown). The activities that may be supported by AI models or tools are coloured in black or grey. The model is designed to be easily exported in .html format and embedded in any webpage, as well as in any other format that supports .html integration. As this is the 0.1 version of the interactive sunburst diagram, work is far from finished. Several updates are needed to accurately represent the data and to render the model more user-friendly and functional. Nevertheless, in their current state, both the cross-functional chart and the sunburst diagram begin to show the complexity of digitization and the areas in which AI can be engaged during its various processes. We hope to publish the final version of the interactive sunburst model on the InterPARES AI website for public access.



Metadata management
Metadata gathering
Metadata models
Metadata enrichment
Cataloguing and description
Definition of data types and values
Digital object identifiers
Metadata about relationships
Metadata creation
Metadata validation/quality control
Digital infrastructure identification and preparation
Software needs and licensing
Security needs
Network needs
Backups and redundancy plan for everyday business
Disaster recovery plan for business continuity in emergencies
Digitization equipment requirements
Decision on external or internal digitization
File storage needs
Final destination/Trusted digital repository
Security assessment
File storage
During and after digitization, before transfer to a trusted digital repository
Creation of protocols
Quality control
Quality control of digitization systems
Decision on requirements for quality control and at which points
Physical collections preparation
Removal of physical barriers to digitization
Physical reconditioning and rehousing
Cleaning
Digital preservation
Integrity checks
Format conversion
Format migration
Communication and outreach
Promotion of digitized collections including to targeted audiences
Partnerships
Management/publication of digitized collections in electronic system
File transfer
Format conversion
Integrity checks
Format migration
Ingestion into trusted digital repository
Access and use management
Sensitivity/security classification
Rights management
Access policy
Copyright and licensing
Physical collections selection for digitization
Selection based on criteria
Risks assessment
Prioritization of materials for digitization
Definition of selection criteria
Condition assessment
Conversion
Text recognition and post-correction
Quality control of digital objects
Capture/creation of digital objects
Colour management - definition of colour profile/standards
Editing for faithful reproduction
Creation of different file versions
Creation of hash value
Creation of technical specifications
Unique ID
Technical parameters
OCR or Non-OCR
Folder structure
File format standards
Cryptographic security decisions
File naming conventions
File versions

Figure 2. First layer of the interactive sunburst model.

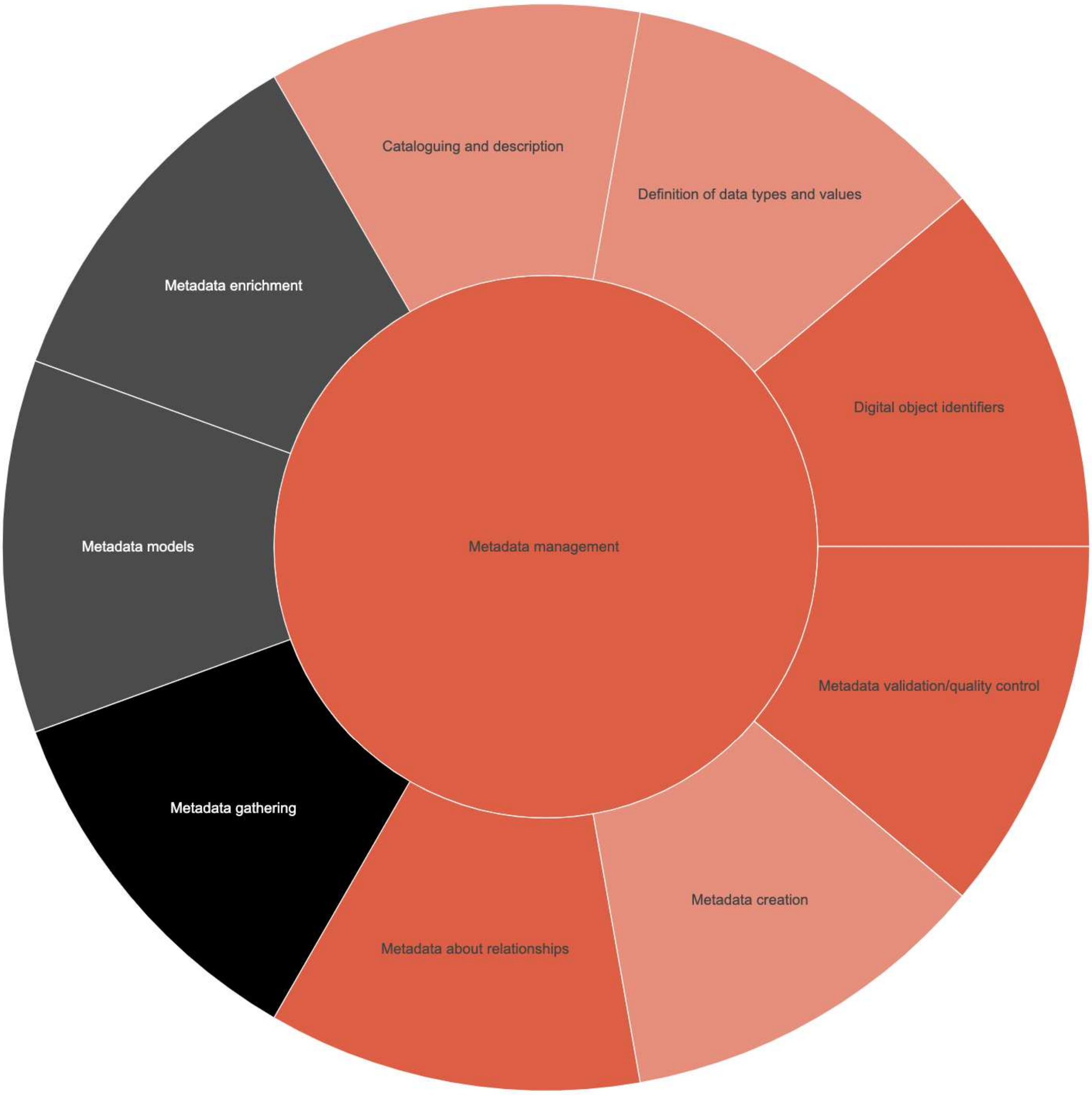

Cataloguing and description
Definition of data types and values
Metadata enrichment
Digital object identifiers
Metadata models
Metadata management
Metadata validation/quality control
Metadata gathering
Metadata creation
Metadata about relationships

Figure 3. Second layer of the interactive sunburst model.

Future work

As organisations continue to digitise their archives, the question of how digitization processes may be improved through the intervention of AI technologies remains an important one to explore. Beyond this, questions around transparency, accuracy, fairness, privacy, and other ethical concerns when using AI - like the production of descriptions of digitised collections, such as video footage, audio recordings, photographs and documents - are equally important to consider. The visual representation of digitization phases and activities and potential AI-supported tasks, as conveyed in our models for AI-assisted digitization, capture a moment in time, when changes in the spectrum of digitization processes are being introduced with the popularisation of AI tools. The forthcoming results of our survey, exploring how organisations are currently using AI during archival digitization processes, will further enrich this picture, and - we hope - contribute to a larger discussion around the impact of AI in the digital representation of, and access to, archives and documentary heritage materials.

References

Federal Agencies Digital Initiatives Guidelines (FADGI), Still Image Working Group (2023). *Technical Guidelines for Digitizing Cultural Heritage Materials: Third Edition (2023 Revised Guidelines)*. Accessed 14 March 2024: https://www.digitizationguidelines.gov/guidelines/digitize-technical.html.

International Organization for Standardization (2010). *ISO/TR 13028:2010 - Information and documentation - Implementation guidelines for digitization of records.*

Plotly. Accessed 14 March 2024: https://plotly.com/

Eng Sengsavang is Reference Archivist at UNESCO Archives. She is co-editor with Jens Boel of Recordkeeping in International Organizations: Archives in Transition in Digital, Networked Environments *(Routledge, 2021).*

Željko Trbušić is a teaching assistant at the University of Zagreb (Croatia), Faculty of Humanities and Social Sciences. He received his PhD in the field of Information and Communication Sciences in 2022.

Annotation of Digitised Archival Materials Supported by AI

by Hrvoje Stančić and Željko Trbušić

Introduction

It is a well-known fact that archives are facing challenges in dealing with digital materials. On the one hand, the challenges stem from the overwhelming number of digital records pouring into the archives, and therefore need to be approached using big data principles grounded in archival theory. On the other hand, fast technological advancements make file formats and media obsolete, to the point that archivists specialising in digital preservation constantly need to assess the influence of new technologies on the records' authenticity, integrity, reliability, and usability. The difficulties often span across several areas, e.g. ICT infrastructure, legislation, finance, education etc. (Mosweu and Bwalya, 2022).

Artificial intelligence (AI) is not a new thing. It was mentioned as early as the first half of the 20th century (McCulloch and Pitts, 1943), and subsequently advanced to the form of expert systems and similar, specialized solutions. Nowadays, the earlier forms of AI are sometimes referred to as "the old AI". To create an AI-based solution, one needs to train it, i.e. provide it with a large enough set of reference materials, and, to oversimplify it, let it learn. Training sets may contain bias and privacy information so one should be careful to avoid them.

The archives, having large numbers of records – either analogue ready to be digitised, or already in digital form – represent a wealth of possible AI training materials. In the case of bias, we might argue that it would always be included in the historic archival materials. And one should not correct it unless the AI trained on them is used for other than understanding the past. However, it is of the utmost importance that the archives that has trained an AI tool on such materials makes it clear that the materials on which AI is trained includes biases.

AI experts may benefit from using archival materials as a training set because they are always on the lookout for new materials to train AI and refine it. Archives may also benefit from cooperation with the AI experts in developing AI solutions that can speed up or automate some of the tedious and repetitive archival tasks. But what are those, and which challenges are critical and can be addressed by AI? We have tried to find answers to those questions.

Identification of critical archival challenges

In the InterPARES Trust AI project, the study "Identification of critical archival challenges that are the best candidates for improvement using AI technologies in the context of retention and preservation of digital records" first conducted a global online survey to respond to the research question posed in its title, and then followed up with a series of in-person interviews. Although narrower in the scope, i.e. focusing on the challenges related to retention and preservation, the study achieved results relevant to all archival functions. Only the results important for this paper are discussed briefly.

The study, conducted simultaneously in English, Spanish, and Portuguese, received a total of 106 responses from 27 countries. The responses came from government archives (50%), college or university archives (18%), corporate archives (3%), international organization archives (3%), special collections (3%), museums (2%), and other cultural institutions (17%).

To the question of whether they have any processes which can be improved by AI technologies, 59% responded positively. *Figure 1* shows to which group of activities the identified processes relate to. When asked whether any of the digital preservation processes in their institutions/organizations involve repetitive or time-consuming tasks, 10% of the respondents confirmed that they have repetitive tasks, 13% have time-consuming tasks, and 30% have both, which results in 61% of the respondents having either one or the other, or both. This clearly represents a good potential for implementation of AI technology.

When asked to identify those repetitive and/or time-consuming tasks, the respondents have included adding, gathering, and extracting metadata as the most common, closely followed by the process of digitization (in general), and processes related to capturing of records. This was confirmed during the in-person follow-up interviews, where the respondents, when asked to explain how they think AI might help them solving their records and archival issues, agreed that AI can help with transcription, acquisition, description, classification, etc. Those responses prompted the next phase of the research during which we have defined an easy-to-follow workflow for training AI to help archives with the creation of image descriptions.

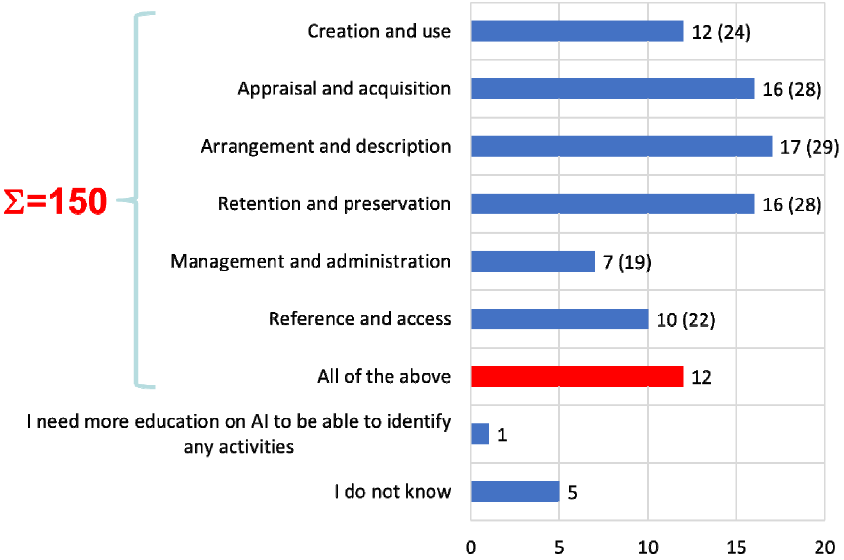


Figure 1. To which group of activities the identified processes which can be improved by AI-related technologies best relate to (multiple answers allowed) (n=63)? Image by authors.

AI workflow

The aim of the second phase of the research was, on the basis of the results of the survey and interviews, defining an AI training workflow with two possible results. The first result of such a workflow would be an AI solution trained on the archival materials ready to be used by archivists who need such a solution. The second result would be a workflow which any archivist with average IT skills can use to develop his/her own trained AI.

To achieve this, we have established a cooperation with The State Archives in Osijek (DAOS), Croatia. They have provided a collection of (set – studio and outdoor) portraits from 1870s to the beginning of the 20th century (*Figure 2*) along with the description of the set, totalling 1,417 images (708 *recto*; 709 *verso*).

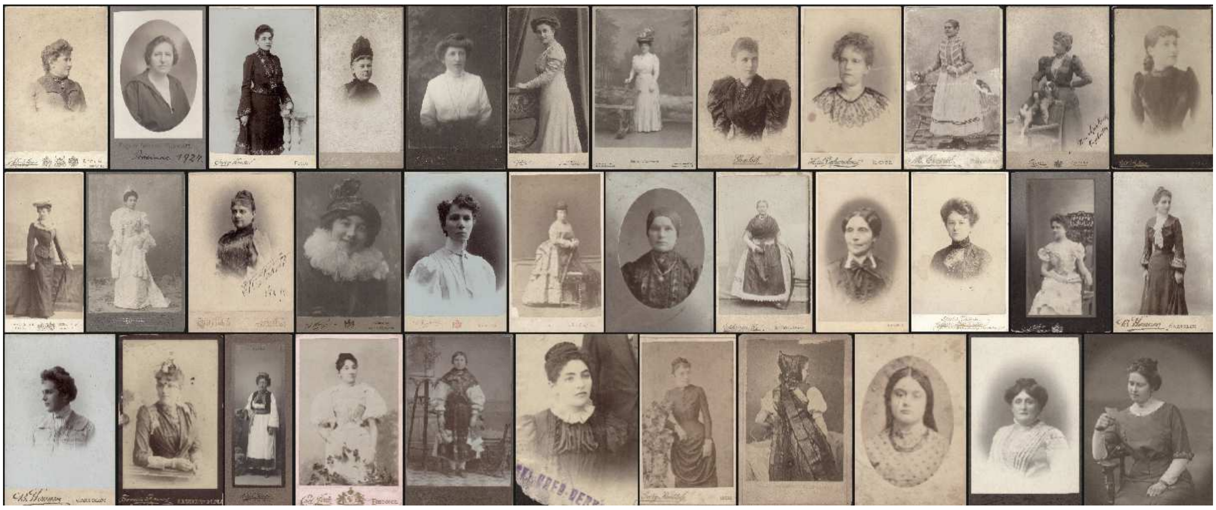


Figure 2. Part of the image collection used for training an AI model. Images from DAOS.

First, the training dataset needed to be prepared. Unique labels were identified and extracted, and this resulted in more than 100 unique labels. For piloting the workflow we have used five (uniform, suit, dress, hat, flowers) – just enough to make sure that the workflow would be functional. For the same reason, we have used 238 images as a training set (roughly one third of the whole set). That set was divided using a 75:25 ratio, i.e. 177 images were used as a training set while the remaining 61 were used as validation set.

Second, to annotate the images using five labels, we have used the Make Sense annotation tool (*Figure 3*). We have labelled all 238 images because the validation set is used by the AI training algorithm to internally validate its training. Once done, the annotated set was exported to the YOLOv5 format.

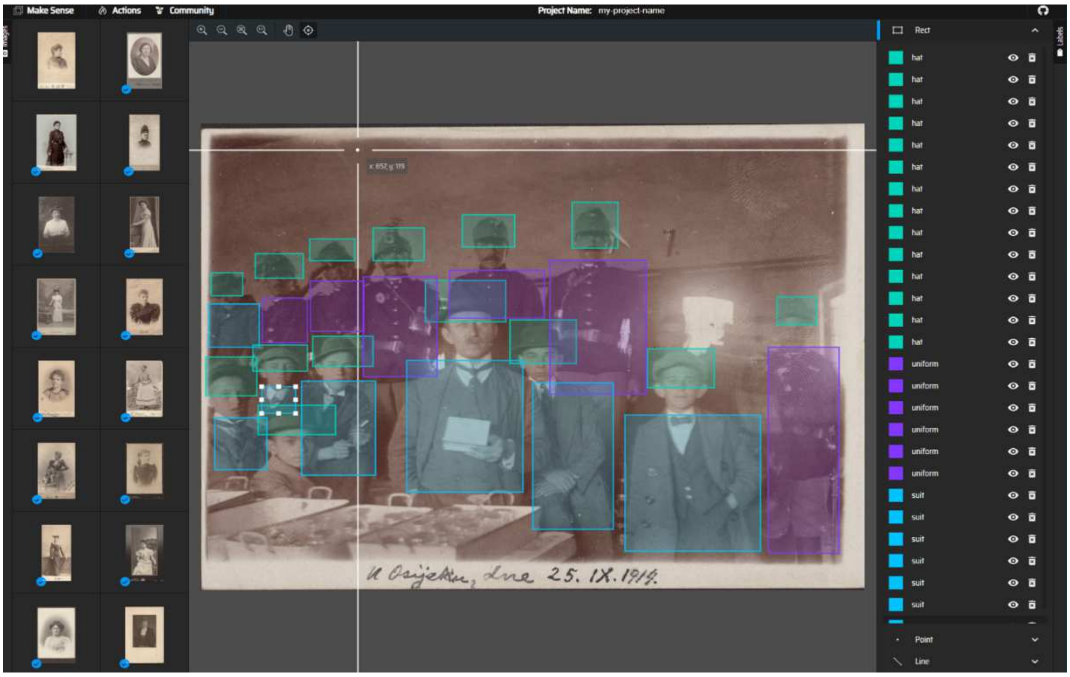


Figure 3. MakeSense annotation tool. Image by authors.

Third, for the AI training environment we have used YOLOv5 (a model in the You Only Look Once [YOLO] family of computer vision models) and PyTorch (open-source machine learning [ML] framework) utilities. We ran AI training in the Google Colab environment, and, given the small size of the training set, after short 21.48 minutes, we had the trained model.

Fourth, we tested the trained model on images from the same collection that were not used for development of the model, on images from other available collections, on images from a similar time-period found on the web and on our own mobile phone-made photographs of the historic images (*Figure 4*).

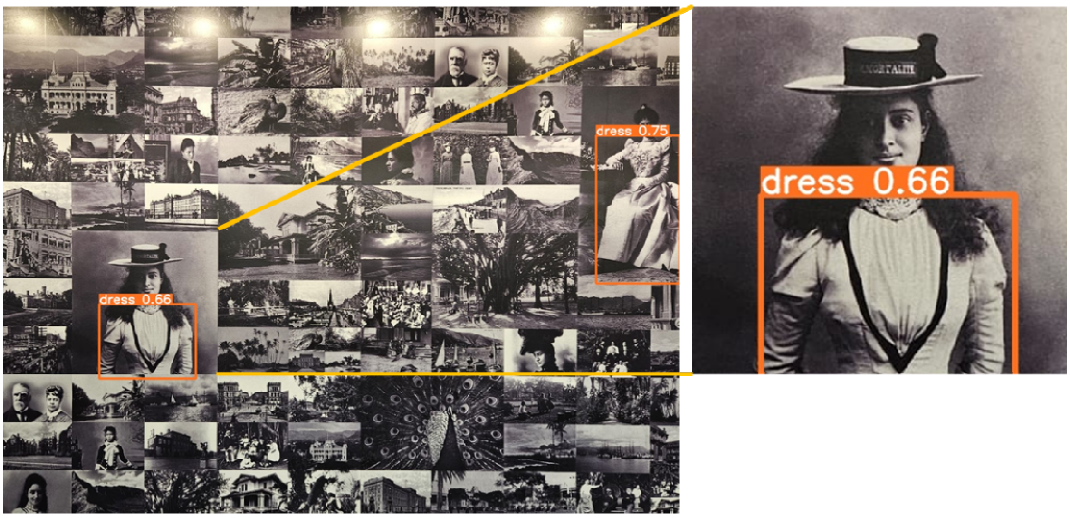


Figure 4. A mobile phone photograph of a reception wallpaper at the Sheraton Princess Kaiulani hotel in Honolulu, Oahu, Hawai'i, USA (left), and the result of the AI model recognition (right). Photo by H. Stančić.

Discussion

It can be concluded that the model was surprisingly successful in the recognition of the trained objects on unlabelled images, given its size. It was also, expectedly, performing poorly in the recognition of flowers in the unlabelled images, as shown by the precision-recall curve for the flowers label (*Figure 5*, left), since the training set contained only few images with flowers, as can be seen from the graph showing the number of label instances in the training set (*Figure 5*, right).

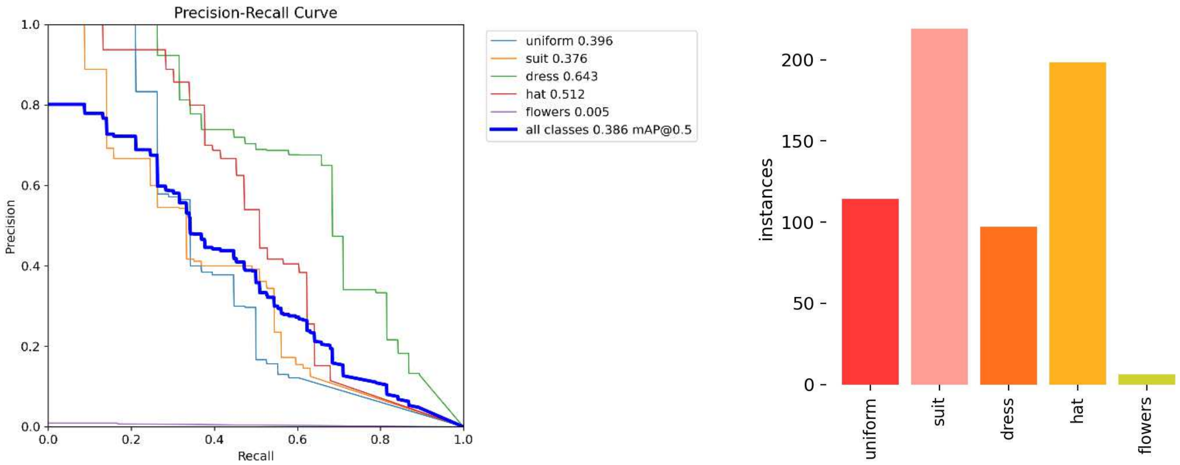


Figure 5. Precision-recall graph (left), number of label instances in the training set (right). Images by authors.

Conclusion and future work

The defined workflow is a result of many tests and errors. There are many solutions that can be used to annotate images, to train AI models for finding the elements in the images, for browsing the set that was used for training, etc., and it was not so straightforward to identify solutions which work well in concert and are at the same time (relatively) easy to use.

The next steps will be to perfect the model on the full set of images using all 100+ labels. After that, the model could be further refined by increasing the number of training images. Since there is a fixed number of images in the collection, additional images could be generated in a way that the original images are a little bit rotated to the left, then to the right, possibly using different angle rotation combination, than further edited to be a little bit darker, or lighter, etc. In this way the training set can be artificially enlarged. Possibly, ChatGPT could be used to automate the process of generation of such a set.

Once the trained AI model will be perfected and refined, we will setup a server environment with the AI model trained on the archival collection of portraits. This will allow any archives to label their collections of digitized portraits by using the model trained on archival images on their set of archival images. This will also allow anyone to effectively reduce the time needed for the repetitive and time-consuming process of "adding, gathering, and extracting metadata". Finally, the workflow itself can be used by any archives to repeat the process by setting up their own training environment and creating AI models to help decrease the amount of repetitive and time-consuming tasks.

References

Make Sense. Available at: https://www.makesense.ai/ (accessed March 11, 2024).

McCulloch, W., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. In *Bulletin of Mathematical Biophysics 5*.

Mosweu, O., and Bwalya, K. J. (2022). The challenges of post custodial management of digital records in Botswana laid bare. In *Information Development*. Available at: https://doi.org/10.1177/02666669221114867 (accessed March 11, 2024).

PyTorch. Available at: https://pytorch.org/ (accessed March 11, 2024).

YOLOv5. Available at: https://github.com/ultralytics/yolov5 accessed March 11, 2024).

Dr. Hrvoje Stančić is Vice-dean for organization and development, and full professor at the Faculty of Humanities and Social Sciences, University of Zagreb (Croatia), where he teaches in the Department of Information and Communication Sciences at undergraduate, graduate and postgraduate levels. In the context of the 4th InterPARES project (2013-2019) he was Director of the European research team where he led a blockchain-related research study. At the Croatian Standards Institute, he is President of the mirror technical committee for development of ISO/TC 307 Blockchain and Distributed Ledger Technologies standard.

Željko Trbušić is a teaching assistant at the University of Zagreb (Croatia), Faculty of Humanities and Social Sciences. He received his PhD in the field of Information and Communication Sciences in 2022.

AI Culture and Images

by Jessica Bushey

Introduction

The integration of AI technologies into the creation of photorealistic images as well as in archival practices aimed to enhance access to and use of digital images archives (both digitized and born digital) has significant implications for the trustworthiness of images as public records and images archives as historical evidence. This essay introduces the concept of AI culture in the context of digital images, one in which technology reshapes traditional practices and challenges established norms. At this stage of research, it is useful to explore the intersection between AI, visual culture and archives to better understand their influences and development. Two *I Trust AI* studies are relevant to such purpose, specifically the "Recordkeeping practices of creators using AI to generate images" and "Increasing access to photos, videos and social media records through AI-generated descriptive metadata". This article will discuss the early findings of the study on "Recordkeeping practices of creators using AI to generate images".

Generative AI

The emergence of generative AI, a type or subset of AI trained to produce new content, either randomly or based on prompts provided by users (McKinsey and Company, 2023), is restructuring productivity across various sectors, from image creation to data analysis (Djanegara et al., 2024). Generative AI tools include large language models (LLMs) such as ChatGPT and image generators such as DALL-E, Midjourney and Adobe Firefly. These tools are now being used to create and curate digital records. The activities of creation and curation are not traditionally associated with the role of the archivist; however, the nature of born-digital records' inherent preservation challenges, such as technological obsolescence, and access challenges, such as volume and complexity, requires intervention by experts in recordkeeping. Prior research products of the InterPARES 2 Project, the "Creator Guidelines" and the "Preserver Guidelines" (http://www.interpares.org/ip2/ip2_products.cfm) offer recommendations for guiding individuals and organizations in creating and preserving trustworthy digital records. The recent proliferation of generative AI models and their application in creating images and curating digital images archives presents an opportunity to revisit these guidelines and exploring the impact of integrating AI technologies into imaging workflows and archival practices. At this early stage of AI culture, as the use of generative AI becomes more widespread in the public and private sectors, experimental and playful applications of synthetic images have emerged alongside intentionally deceptive ones, igniting debates over equating seeing with believing (Hsu and Myers, 2023). "As AI-generated content becomes more prevalent and difficult to distinguish from human-generated content, individuals may become more skeptical and distrustful of the information they receive"

(Djanegara et al., 2024, p. 11). AI culture places the very notion of trustworthy records in peril. As stewards of public records and historical archives, we need to make informed decisions prior to the acquisition of generative-AI images and the adoption of generative-AI tools for archival practices.

Literature Review

A review of the literature and industry-led initiatives addressing AI-generated images revealed a lack of contributions from archival scholars and professionals, in particular of a discussion aimed to an understanding of current recordkeeping practices for creation and use of synthetic images as an emergent record format. On the other hand, an increasing number of articles in academic and professional archival journals have explored automation and the application of AI to the field of cultural heritage and recordkeeping (Lee, 2018; Hedges et al., 2022; Jaillant and Rees, 2023; Bunn, 2020). In a survey of literature on archives and AI, conducted by Colavizza et al. (2021), the authors discuss the growing use of AI technologies and tools in archival processes, with the aim of automating aspects of archival appraisal, digitization, metadata creation and extraction, and providing access to archives for a more diverse range of users. Understandably, archivists' initial interest in AI is to solve archival challenges. But, as individuals and organizations gain skills to integrate AI applications into workflows, we can anticipate exponential growth in the circulation of AI-generated images created and disseminated as records in fields that involve public trust. Records managers and archivists will soon need to be familiar with AI technology and tools to inform decisions and actions taken when acquiring born-digital images and providing access to archival images.

As part of the I Trust AI study entitled "Recordkeeping practices of creators using AI to generate images", a review of literature on AI-generated images identified several fields that discuss the topic in the context of verification of authenticity, concerns about accuracy and reliability, and fake image detection. They are medicine, law enforcement and journalism and media communications (Bushey, 2023). The applications of AI technology and tools in these fields vary, yet reoccurring themes were identified throughout the literature:

(1) Authenticity and Verifiability – approaches and activities focusing on contextual information that contributes to the creation and use of AI-generated images as trustworthy records.
(2) Manipulation and Misinformation – approaches and activities contributing to identifying images that have been intentionally altered or created as fakes.
(3) Bias and Representation – approaches and activities that focus on training datasets for generative AI.
(4) Attribution and Intellectual Property – approaches and activities specific to the rights and responsibilities of AI-generated images.

(5) Transparency and Explainability – concerns specific to the proprietary nature of AI technologies and tools.
(6) Ethical Considerations – concerns and approaches specific to privacy requirements (Bushey, 2023).

Research Questions

The literature review provides a current snapshot of what is known about AI-generated images and the application of AI technology and tools in the fields of medicine, law enforcement and journalism and media communications. A discussion of the thematic areas reveals shared concerns and industry-led initiatives that both inform archival research on generative-AI and invite contributions from archival experts. It also led to the identification of research questions to guide the next phase of the study on "Recordkeeping practices of creators using AI to generate images":

(1) How are individuals and organizations using AI-generated images?
(2) What AI tools and technologies are being used to create, manage, and store AI-generated images?
(3) What actions are being taken by individuals and organizations to identify AI-generated content?
(4) What standards and/or policies are guiding procedures for creating, using, and preserving AI generated images? (Bushey, 2023).

Conclusions

These research questions will guide the next phase of the study, which will continue to explore the intersection between AI, visual culture, and archives. When drawing on literature and initiatives from other disciplines we can see that AI technologies and tools are disrupting established procedures for creating, using, and managing images. If we approach AI-generated images as an emergent record format, it presents an opportunity to analyze its nature and characteristics. Adobe's recent media-focused Content Authenticity Initiative (CAI) and the 2019-founded Project Origin, a collaboration between BBC and Microsoft focusing on digital journalism, assert the importance of digital image metadata that capture both content and context specific to generative-AI (Bushey, 2023). As happened in the past, when digital imaging disrupted analogue photography, this situation presents an opportunity to revisit concepts of provenance and trustworthiness in the AI era. Archivists are well positioned to conduct this type of research through interdisciplinary collaboration.

References

Bunn, J. (2020), "Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI)", *Records Management Journal*, Vol. 30, No. 2, pp. 143-153. https://doi-org.libaccess.sjlibrary.org/10.1108/RMJ-08-2019-0038

Bushey, J. (2023). “AI-Generated Images as an Emergent Record Format.” *023 IEEE International Conference on Big Data (BigData)*. Sorrento, Italy, 2023. pp. 2020-2031. doi: 10.1109/BigData59044.2023.10386946.

Colavizza, G., Blanke, T., Jeurgens, C., and Noordegraaf, J. (2021). “Archives and AI: An Overview of Current Debates and Future Perspectives,” *Journal on Computing and Cultural Heritage*, vol. 15, no. 1, pp. 1–15. doi: 10.1145/3479010.

Dewi Toft Djanegara, N., Zhang, D., Badi Uz Zaman, H., Meinhardt, C., Watkins, G., Nwankwo, E., Wald, R., Kosoglu, R., Koyejo, S., and Elam, M. (February 2024). “Exploring the Impact of AI on Black Americans: Considerations for the congressional Black Caucus’s Policy Initiatives.” White paper, v1.0. Stanford Institute for Human-Centered Artificial-Intelligence. https://hai.stanford.edu/sites/default/files/2024-02/Exploring-Impact-AI-Black-Americans.pdf

Hedges, M., Marciano, R., and Goudarouli, E. (2022). “Introduction to the Special Issue on Computational Archival Science.” *Journal of Computing & Cultural Heritage*, vol. 15, no. 1, p. 1:1-1:2. doi: 10.1145/3495004.

Jaillant, L., and Rees, A. (2023). “Applying AI to digital archives: trust, collaboration and shared professional ethics.” *Digital Scholarship in the Humanities*, vol. 38, no. 2, pp. 571–585. doi: 10.1093/llc/fqac073.

Lee, C. (2018). “Computer-assisted appraisal and selection of archival materials.” *IEEE International Conference on Big Data (Big Data)*. 2018. pp. 2721–2724. doi: 10.1109/BigData.2018.8622267.

McKinsey and Company (2023, January 19). “What is Generative AI?” Available via https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai.

Hsu, T., and Myers, S. L. (2023, April 8). Can We No Longer Believe Anything We See? *The New York Times*.
https://www.nytimes.com/2023/04/08/business/media/aigenerated-images.html

InterPARES Trust AI (2021-2026). "Research Studies." Available via https://interparestrustai.org/trust/about_research/studies

Xing, X., Terhörst, P., Raja K., and Pedersen, M. (2024) "Analyzing Fairness in Deepfake Detection with Massively Annotated Databases." Version 4. Available via arXiv:2208.05845v4

Dr. Jessica Bushey is an Assistant Professor in the School of Information at San José State University in California (USA), where she teaches courses on Reference and Information Services in Archives, and Preservation Management in Archival Repositories. Prior to joining SJSU, Bushey worked with municipal and university museums and archives, and international organizations to develop policies and procedures for managing and preserving digital images and audiovisual collections.

The Africa-Europe Gaps of AI Regulations for Managing Public Records

by Esteban Guerrero, Proscovia Svärd, Tolu Balugon, Nampombe Saurombe, Pekka Hentonnen, and Lorette Jacobs

Introduction

Public records management is the practice and research area focused on creating and maintaining the official records generated by a government agency (Gabriel, 2008). It involves the lifecycle of a record, from its creation to its disposition (destruction or preservation). Therefore, public records management and *e-government* practices are intrinsically linked as a foundation for the delivery of e-services, and public records are the *raw materials* for many e-government services.

With the recent introduction of Artificial Intelligence (AI)-based systems in almost every aspect of the society, *electronic records* and their management (classification, storage, retrieval and disposition) and AI developments are tightly bound. This article therefore focuses on the impact of autonomous software on the management of public records in the context of e-government. There are four public records management tasks that AI could contribute to:

1) Enhanced efficiency and organization: AI can automate tasks like classifying, indexing, and tagging public records ((Al-Mushayt, 2019; Jovanovic et al., 2014; Rolan et al., 2019), which would free records managers'/archivists' time for more complex tasks and ensure consistency in record-keeping.
2) Improved search and retrieval: AI may analyse large amounts of public records and identify relevant information based on user queries (Baron et al., 2022; Masenya, 2020), which may allow citizens and government officials/institutions to find the information they need more quickly and easily.
3) Transparency and accountability: AI can be used to identify and redact sensitive information in public records before they are released, ensuring transparency while protecting privacy (Baron and Payne, 2017; Hofman, 2020; Ingram and Johnson, 2022; Modiba et al., 2019). Additionally, AI may track the use of public records (Anantrasirichai and Bull, 2022; Zhang et al., 2021), making it easier to hold government agencies accountable for their actions.
4) Data insights and innovation: AI-based mechanisms could be used to analyse public records identifying trends and patterns, providing valuable insights for policymakers and service providers. This data can then be used to develop new and improved e-government services.

However, a literature review (Svard et al., 2024, forthcoming) reveals a critical disconnect between existing *AI regulations* and public records management in Sweden, Finland, and South Africa. National guidelines of these countries lack clear specifications on appropriate technologies for handling public records. Notably, none of the reviewed regulations mentions public records management, focusing only on data management and innovation. This is a significant oversight, considering the potential for opacity in algorithms and automated decision-making processes used in creating e-services. Even though open data initiatives incorporate public records, the effective management of the records behind these algorithms remains crucial.

The European advantage and ongoing developments

The European Commission's approach to embedding AI regulations within the legal framework of EU countries is a promising step, which is focused on establishing an EU citizens-oriented legislation for the application of AI-based systems to every aspect of the EU society, including cyberspace (European Commission, 2021). These efforts are being emulated by South Africa and the African Union as a whole (South African Government, 2021). The authors highlight potential opportunities to bridge the gap between AI regulations and public records management at an African and European level. This interdisciplinary work suggests that the European framework for AI regulation stands out as the most comprehensive to date. It defines four risk levels (unacceptable, high, limited, minimal) and sets clear requirements and obligations for AI systems, actors, and governance structures. In contrast, Africa lacks a unified approach. While various initiatives exist at national, regional, and continental levels to govern AI development, a recent study by Daigle (2021) reveals that only 14 out of 55 African countries possess data protection legislation addressing automated decision-making, a core aspect of AI. These academic efforts by Svard et al. and Daigle highlight the need for a more harmonized approach to AI regulations in Africa that considers public records management within its framework.

AI regulations as indicators of national priorities:

The lack of public records management considerations in AI regulations suggests the need to involve information professionals, especially those in the records and archives management field, in the formulation of such regulations. The accountability and transparency of government institutions hinges on the effective creation and capture of the records that accrue out of processes where AI/automated decision-making have been deployed.

The authors are conducting a case study as part of I Trust AI in Sweden, Finland and South Africa where local authorities and government organizations are being asked the following questions:

RQ. 1. What are the areas where automated decision-making/AI is being deployed?
RQ. 2. What are the records management challenges?

Preliminary results from the Swedish case suggest that the Swedish government has implemented several e-services, and government organizations have rolled out a variety of systems (including AI-based) that have transformed how they serve the Swedish citizens. These systems include complex e-services and automated processes for routine tasks, especially for handling large volumes of cases. This process has to some extent freed civil servants from monotonous work to focus on more creative work, but the management of the records that grow out of these processes is not clear.

The Finnish case shows that legislation sets limits to the usage of AI in decision-making processes articulating criteria that must be met if AI is to be used in decision-making. Automated decision-making is only allowed if case-by-case consideration is not needed. In addition, there must be documentation of how a decision was reached by officials who bear responsibility for it. In case of any errors, it should be possible to identify them.

South Africa has witnessed the adoption and deployment of automated decision-making in various sectors driven by advancements in technology and the pursuit of efficiency. One noteworthy application of automated decision-making pertains to the financial sector. Some banks in South Africa make use of automated decision making to provide a customer profile and this might include customer performance at his/her place of work, credit worthiness, location, health, reliability, personal preferences, or conduct. However, the code of conduct of the Banking Association of South Africa requires that the banks ensure that the customers whose data is collected in this manner, are informed of their rights in respect of automated decision making. This means that data subjects should have the ability to make representations about an automated decision as well as make use of complaint procedures. The information from the South African private sector is regarded as public and should be made accessible to the citizens.

These preliminary results of our research demonstrate the need for effective public records management regimes to capture, manage and retrieve the records required to promote transparency and accountability in processes where AI/automated decision-making is being applied by government agencies. For this, collaboration between records managers and archivists and the professions implementing AI/automated decision making is necessary.

Recommendations for building a solid foundation for AI governance: research, policy, and collaboration

The academic literature underscores the need for a multi-disciplinary approach to address the integration of AI with public records management. We propose a set of recommendations for building a sturdy foundation for AI governance where research, policy, and collaboration meet:

1) **Prioritizing robust records management:** Future research and policy development should place importance on *robust* public record management

practices alongside AI implementations. This will ensure transparency and accountability in government decision-making processes.

2) **Fostering international and regional collaboration:** We advocate for collaborative initiatives between European and African countries where government authorities and records management and AI practitioners could find an agreement on:
    a. *Skills development*: Equipping professionals with the necessary skills to navigate AI governance.
    b. *Knowledge exchange*: Sharing best practices in building responsible AI frameworks for records management.
    c. *Dissemination of best practices*: Making successful strategies accessible to all stakeholders.
3) **Harmonized frameworks in Africa**: We recommend the creation of a unified AI regulatory framework across African nations, based on the European Union's model. This framework should prioritize comprehensive legislation that addresses the following African needs:
    a. *Data protection*: Ensuring the responsible handling of citizen data.
    b. *Automated decision-making*: Establishing clear guidelines for AI-driven decision-making processes.
    c. *Ethical considerations*: Integrating ethical principles into the development and use of AI.
4) **Building an AI-Ready workforce**: Investing in education, training, and skill development is crucial to create a workforce equipped to manage the challenges and opportunities of AI governance and innovation. This skilled talent pool will be essential for responsible AI implementation.
5) **Data-driven policymaking**: Regularly collecting and analysing data on regulatory advancements and setbacks. This will provide valuable insights into the effectiveness of AI governance policies. Utilizing these metrics as guiding principles will ensure that policymaking remains evidence-based and responsive to societal needs.

Conclusion

This paper offers a critical perspective on AI legislation and its impact on public records management. By adopting the proposed recommendations, stakeholders can lay the groundwork for a future where AI serves to enhance public services, uphold democratic values, and promote sustainable development. This will ensure responsible AI governance that aligns with societal expectations.

References

Al-Mushayt, O.S. (2019). Automating E-government services with artificial intelligence. *IEEE Access* 7, 146821–146829.

Anantrasirichai, N., and Bull, D. (2022). Artificial intelligence in the creative industries: a review. *Artificial Intelligence Review* 55, 589–656. https://doi.org/10.1007/s10462-021-10039-7.

Baron, J.R., and Payne, N. (2017). Dark archives and edemocracy: strategies for overcoming access barriers to the public record archives of the future. In *2017 Conference for E-Democracy and Open Government (CeDEM)*. IEEE, pp. 3–11.

Baron, J.R., Sayed, M.F., Oard, D.W. (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 of Computation & Cultural Heritage* 15, 5:1-5:19. https://doi.org/10.1145/3481045.

Daigle, B. (2021). Data Protection Laws in Africa: A Pan-African Survey and Noted Trends. *Journal of International Commerce & Economics* 2021, 1.

European Commission (2021). Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Brussels.

Gabriel, S. (2008). Public Sector Records Management: A Practical Guide. *Records Management Journal* 18. https://doi.org/10.1108/rmj.2008.28118bae.001.

Hofman, D. (2020). "Between knowing and not knowing": privacy, transparency and digital records (PhD Thesis). University of British Columbia.

Ingram, W.A., and Johnson, S.A. (2022). *Ensuring Scholarly Access to Government Archives and Records.*

Jovanovic, J., Bagheri, E., Cuzzola, J., Gasevic, D., Jeremic, Z., Bashash, R., (2014). Automated semantic tagging of textual content. *IT Professional* 16, 38–46.

Masenya, T.M. (2020). Application of modern technologies in the management of records in public libraries. *Journal of South African Society of Archivists* 53, 65–79.

Modiba, T., Ngoepe, M., Ngulube, P. (2019). Application of Disruptive Technologies to the Management and Preservation of Records. *Mousaion South African Journal of Information Studies* 37. https://doi.org/10.25159/2663-659X/6159.

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., Stuart, K. (2019). More human than human? Artificial intelligence in the archive. *Archives & Manuscripts* 47, 179–203.

Smith, K. (2016). *Public Sector Records Management: A Practical Guide*. Routledge, London. https://doi.org/10.4324/9781315602998.

South African Government (2021). Minister Khumbudzo Ntshavheni: Artificial intelligence regulation while encouraging innovation | South African Government.

Svard, P., Esteban, G, Tolu, B., Saurombe, N., Lorette, J., Henttonen, P. (2023). Local regulations for the use of artificial intelligence in the management of public records – A literature review. Submitted to the *Records Management Journal*.

Zhang, Q., Lu, J., Jin, Y. (2021). Artificial intelligence in recommender systems. *Complex Intelligent Systems* 7, 439–457. https://doi.org/10.1007/s40747-020-00212-w

Proscovia Svard is an Associate Professor at the Department of History, Sorbonne University, Abu Dhabi, UAE. She teaches Records Management and Archival Science.

Esteban Guerrero is an Associate Professor at Umeå University, Sweden. His work focuses on formal methods of software agents in Artificial Intelligence.

Tolulope Balogun is a Postdoctoral researcher at the University of South Africa with interest in digitization, digital preservation, Artificial Intelligence, records and archives.

Nampombe Saurombe is a Professor at the Department of Information Science, University of South Africa in Pretoria, South Africa. She is also the co-chair of the International Council on Archives' (ICA) expert group on research and outreach services (EGRSO).

Pekka Henttonen is an Associate Professor, Faculty of Information Technology and Communication Sciences of Tampere University, Finland. His focus in research has been in recordkeeping metadata and archival knowledge organization.

Lorette Jacobs is a Professor and the Chair of the Department of Information Science at UNISA, Pretoria, South Africa. She is currently involved in the InterPARES Trust AI project.

AI Literacy: A Must for Records Management and Archival Professionals

by Moises Rockembach

Introduction

Several InterPARES Trust AI studies are notable for their innovative use of AI in archival practices. These include projects designed to transform emergency services with AI-powered data analysis, potentially changing how we respond to public safety issues. Others focus on digitally preserving and analyzing historical scrolls, using deep learning to access and safeguard our shared history. The Digital Twin study, which creates digital versions of physical objects, shows AI's ability to manage and enhance complex systems (Frontoni et al., 2022) These are just a few examples among almost 50 studies carried out by InterPARES Trust AI researchers.

Among the ongoing studies, one investigates the nuances of AI literacy among records management and archival professionals, categorizing proficiency levels from beginners to experts. At the beginner level, individuals gain a basic understanding of AI concepts along with basic knowledge of records management and archiving principles. This phase is crucial to building a solid foundation on which more complex skills can be developed. The intermediate level sees professionals gaining a comprehensive understanding of both fields, including AI methodologies and records management practices and theories. It is at this stage that the combination of traditional knowledge with new technologies becomes more pronounced, leading to innovative approaches to information management.

The Convergence between AI and Records and Archives Management

The intersection of Artificial Intelligence (AI) with Records and Archives Management represents a fundamental evolution in information management. This convergence is driven by the transformative potential of AI technologies to improve the accessibility and preservation of records and archival materials. As we navigate the digital age, integrating AI into these fields is not just an option, but a necessity.

Some authors emphasize the importance of approaching the Data Lifecycle comprehensively—from acquisition, cleansing, and modeling to implementation, optimization, analysis, visualization, evaluation, sharing, erasing, and archiving (Grillenberger and Romeike, 2017). This holistic view is essential for effectively integrating AI into records management and archival workflows, ensuring that information is not only preserved and accessible but also intelligently managed throughout its lifecycle.

AI's role in records and archives management is multifaceted, addressing challenges such as data overload, preservation of digital formats, and the need for more sophisticated information retrieval systems. The adoption of AI technologies offers promising solutions, from automating classification and metadata generation to enabling more nuanced search capabilities through natural language processing (NLP) and machine learning algorithms.

The exploration of AI's application in information science, particularly within the realms of records management and archives, has been steadily growing. Floridi (2014) conceptualizes the potential of this context as part of the Fourth Revolution, wherein information technology, including AI, reshapes our understanding of information's role in society. This revolution positions AI as a critical component in managing the infosphere, where human and computational agents interact in unprecedented ways. The evolution of AI from theoretical constructs to practical applications within archives and records management underscores a shift towards intelligent information environments (Floridi, 2014).

Furthermore, several studies have explored professionals' opinions on AI technologies (Cushing and Osti, 2023), emphasizing the need for ethical approaches to build trust between parties (Jaillant and Rees, 2023), as well as a data culture that integrates AI technologies into archival practices (Melo and Rockembach, 2019; Rockembach, 2021). This involves recognizing the significance of data quality and understanding and leveraging AI's capabilities to improve archival processes and outcomes. AI's contribution to enhancing data quality through advanced processing and analysis techniques signifies a leap forward in how we manage and preserve information over time.

The importance of AI literacy among professionals in this field cannot be underestimated. As Long and Magerko (2020) argue, developing competencies in understanding and applying AI technologies is essential for interdisciplinary collaboration and innovation. In records management and archives, this literacy extends beyond the theoretical aspects of AI, encompassing practical skills in data analysis, algorithmic design, and ethical considerations in AI deployment.

Interdisciplinary research and projects illustrate the practical benefits and challenges of integrating AI in archival and records management practices. For instance, the application of machine learning for document classification and sentiment analysis offers insights into how AI can support more efficient information retrieval and understanding of archival content (Liu and Lee, 2018). Similarly, the use of NLP and computer vision technologies has shown potential in digitizing and indexing historical documents, making them more accessible to researchers and the public (Marciano et al., 2018).

However, integrating AI into the management of records and archives also highlights ethical and operational challenges. Concerns like data privacy, bias in AI algorithms, and the digital divide are significant issues that professionals need to navigate. These ethical implications underscore the importance of developing frameworks and guidelines that ensure responsible use of AI technologies (Rhem, 2021). Addressing these challenges requires a concerted effort by professionals, researchers, and policymakers to create an environment that fosters trust and transparency in AI applications.

Transforming the future for Records Management and Archival Professionals with AI literacy

In order to empower records management and archival professionals with the tools and knowledge necessary for harnessing the potential of AI, it is critical to develop an interdisciplinary AI literacy framework. Based on ongoing research into AI literacy in records management and archives, I will outline the essential skills required for effective participation in projects within these areas. This framework encompasses six pivotal categories: Concepts and Theories, Records and Archives, Computing, AI Ethics, AI User-centric Approach, and AI in the Organization. Each category represents a critical facet of AI literacy, collectively providing a comprehensive understanding and skillset that professionals in the field must possess to navigate and leverage AI technologies effectively.

The first step in developing Archival-AI literacy lies in grounding professionals in general AI **Concepts and Theories**. This foundational category covers the general ideas and theories that underpin AI, including understanding the meaning of machine learning, deep learning, natural language processing, and computer vision. These concepts are the foundation upon which additional AI-related knowledge and applications are built, offering a critical lens through which professionals can evaluate and integrate AI technologies into records and archival management. By understanding these fundamental concepts, professionals can better reason about the capabilities and limitations of AI, preparing themselves for informed decision-making and innovative applications in their respective fields.

The second category, **Records and Archives**, explores how AI intersects with records management and archival practices in key areas. Records literacy equips professionals with crucial skills for the big data era: understanding, handling data, and recognizing patterns. Datafication and digital transformation show the shift to data-driven methods in society and records management. Additionally, knowledge of Identity Metadata and Paradata is essential for maintaining and/or verifying the reliability, accuracy and authenticity of digital records, be they in recordkeeping systems or in archival preservation systems, by tracking the origins and contexts of data.

The third category, **Computing**, is the technical core of AI, vital for applications in records and archives management. Professionals need to understand Machine Learning, including its various forms (supervised, semi-supervised, unsupervised), to use AI for automating data classification, sorting, and retrieval, for example. Computer Vision is key for interpreting visual data, essential in digitizing and preserving archives. Generative AI focuses on creating new content, such as text and images, from large datasets, offering new ways to handle archival materials. Knowledge of Algorithms is also crucial for applying AI effectively to archival processes, ensuring tasks are performed accurately.

The fourth category, **AI Ethics**, is vital for integrating ethical practices in AI within records and archives management. There is a need to address biases and foster diversity in AI projects. AI systems must be ethically sound, prioritizing public benefit, privacy, and rights. *Data Protection* and *sensitivity* aim at securing sensitive data, essential for ethical AI use. *Transparency* and *explainability* stress making AI processes clear and understandable, building trust in AI applications. *Legal concerns* point out the importance of adhering to regulations, like the AI Act in Europe, to ensure AI's legal compliance in archival work.

The fifth category, **AI User-centric Approach**, ensures AI is built and used with a focus on users' needs. It includes *problem-solving with AI*, which uses AI to address complex issues, improving decision-making and efficiency. *AI design* and *interaction* focus on refining AI for better user experiences, making sure systems are user-friendly. *Testing* and *iteration* highlight creating flexible AI that improves with user feedback, maintaining relevance and usability.

The sixth category, **AI in the Organization**, covers integrating AI into institutions with an emphasis on governance and collaboration. *AI adoption* and *governance* are about establishing ethical and effective AI use strategies that match organizational and ethical goals, including setting up policies for its use. *Collaboration* and *communication* encourage teamwork across departments, promoting the sharing of AI knowledge and practices, leading to unified and impactful AI use within the organization.

Conclusion

The ongoing research underscores the rapid growth of AI applications across various sectors, including the records and archives management field. Yet, it also highlights a critical gap: the prevalence of misunderstandings about AI's uses and its capabilities. These misconceptions often stem from a lack of familiarity and engagement with AI technologies, underscoring the importance of AI literacy as a means to demystify AI and unlock its potential responsibly.

It is important to note that, while a deep technical expertise in computing may not be necessary for all records and archival management professionals, a basic knowledge is essential. This ensures they can effectively collaborate in AI projects, communicate with technical teams, and make informed decisions about the use of AI in their work. Regardless of whether they work with analog or digital materials, understanding the possibilities offered by artificial intelligence has become indispensable today. In an increasingly data-driven world, the ability to integrate, analyze, and optimize processes using AI is not just a competitive advantage but a fundamental necessity for the effective preservation, access, and management of information. The capacity to adapt to and employ these innovative technologies will ensure that the field of records and archives remains relevant, agile, and capable of meeting contemporary and future information demands.

The future of AI in records management and archives is contingent on ongoing multidisciplinary collaboration. Building a community of practice that shares knowledge, experiences, and best practices is crucial for advancing AI literacy and capabilities within the field.

References

Cushing, A. L., and Osti, G. (2023). “So how do we balance all of these needs?”: how the concept of AI technology impacts digital archival expertise. *Journal of Documentation,* 79(7), 12-29.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Frontoni, E., Paolanti, M., Lauriault, T. P., Stiber, M., Duranti, L., & Muhammad, A. M. (2022). Trusted Data Forever: Is AI the Answer*?. arXiv preprint arXiv:2203.03712*. https://doi.org/10.48550/arXiv.2203.03712

Grillenberger, A., and Romeike, R. (2017, November). Key concepts of data management: an empirical approach. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research* (pp. 30-39).

Jaillant, L., and Rees, A. (2023). Applying AI to digital archives: trust, collaboration and shared professional ethics. *Digital Scholarship in the Humanities*, 38(2), 571-585.

Long, D., and Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

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 *Re-Envisioning the MLS: Perspectives on the future of library and information science education.* Emerald Publishing Limited.

Melo, J. F., and Rockembach, M. (2019). Archival Science and Information Science in the Big Data Era: Research Perspectives and Professional Practice in Digital Archives. *Prisma.Com: Journal of Information and Communication Sciences and Technologies.* Porto: CIC. Digital. No. 39 (2019), pp. 14-28. https://doi.org/10.21747/16463153/39a2

Rhem, A. J. (2021). AI ethics and its impact on knowledge management. *AI and Ethics*, 1(1), 33-37.

Rockembach, M. (2021). Information Science and Artificial Intelligence: A Path to Smart Archives and Libraries. In *ISKO Spain-Portugal Congress (5th: 2021: Lisbon): Knowledge Organization on the Horizon 2030: Sustainable Development and Health. Proceedings.* [electronic resource]. Lisbon: University of Lisbon, 2021. http://hdl.handle.net/10183/233477

Moises Rockembach is a Professor of Information Science at the University of Coimbra (Portugal) and Researcher at InterPARES Trust AI.

Outlook – What Insights does InterPARES Trust AI Offer Memory of the World?

by Corinne Rogers and Luciana Duranti

The Memory of the World Programme (MoW) was established in 1992 with three objectives:

1. To facilitate preservation, by the most appropriate techniques, of the world's past, present, and future documentary heritage;
2. To assist universal access to documentary heritage;
3. To increase awareness worldwide of the existence and significance of documentary heritage and thereby foster dialogue and mutual understanding between people and cultures.[48]

At that time microfilming and early digitisation projects were being promoted as means of preserving and giving access to fragile analogue media. Documents were increasingly being created in digital form, but their status as records and how to manage them was still being debated. The internet was a niche technology not yet widely and publicly used.

How things have changed! While the goals of the MoW remain constant, the means of achieving those goals continue to emerge and develop with each new iteration of information technology. The MoW has tracked the world's documentary heritage through the MoW Register, and has offered practical assistance and advice to member states through publications, policy development, education and training programmes, exhibitions, prizes and awards, keeping pace with technological advancements. Parallel but separate, InterPARES has since 1998 and through five phases conducted research on the challenges and affordances of developing information technologies for archives and records professionals, with a constant focus on supporting the creation, use, preservation and access of trustworthy records in all types of digital systems. InterPARES has developed knowledge essential to the long-term preservation of the authenticity of records created and/or maintained in digital form, and provided the basis for standards, policies, strategies and plans of action. Like MoW, InterPARES is international in scope, and is supported by an interdisciplinary process that has included a wide range of academic and professional fields, from natural and applied sciences, government and the arts, law, computer science, cybersecurity, and engineering, and now Artificial Intelligence, including machine learning, natural language processing, and deep neural networks.

The five recommendations of MoW map to the current work of InterPARES, as demonstrated through the articles in this issue, and evidenced by many other studies available on the I Trust AI website

[48] UNESCO Executive Board. (2021) General Guidelines of the Memory of the World Programme. 211the Session (211 EX/10 Decision. https://unesdoc.unesco.org/ark:/48223/pf0000378405#:~:text=The%20five%20strategies%20for%20the,and%20national%20and%20international%20cooperation.

(https://interparestrustai.org/trust/about_research/studies). The MoW recommendations are:

1. Identification of documentary heritage;
2. Preservation of documentary heritage;
3. Access to documentary heritage;
4. Policy measures that acknowledge documentary heritage as an invaluable asset, recognized in national legislation, development policies and agendas;
5. National and international cooperation that brings together human and material resources to assist research and protect and preserve documentary heritage, supports the exchange of research data, publications and information, education and training, promotes the organization of meetings and working groups, and encourages cooperation with international and regional professional associations, institutions and organizations.[49]

Identification and preservation of documentary heritage

AI is already being used to identify records requiring preservation by many institutions and organizations. AI tools have long been used by email systems to categorize emails according to their creator's retention requirements, allowing for further automated processing, including bulk destruction or preservation. InterPARES Trust AI is taking this further, exploring means of AI-assisted classification of documents. "Records management will become a data science, overseeing algorithms that apply record classifications and/or record retention and access rules."[50]

The ability to preserve authentic records begins at the time of their creation and is supported through maintenance and handling. Records made and/or processed by AI tools require the inclusion in their contextual information of paradata about the tools themselves, the criteria for adopting them, and the procedures followed for their use and the persons carrying them out.

Access to documentary heritage

AI offers enormous opportunities to provide access to diverse and disparate holdings of material online. While many applaud the ability to search and find content of interest at the touch of keystroke, archivists, records managers, and all conscientious researchers need to be concerned not only with finding material, but with identifying and maintaining the context of that material through associated metadata and paradata that preserve and protect provenance, use, and rights data. This is central to I Trust AI.

[49] UNESCO. (2015) Recommendations concerning the preservation of, and access to, documentary heritage including in digital form. UNESCO, Paris, France. https://www.unesco.org/en/legal-affairs/recommendation-concerning-preservation-and-access-documentary-heritage-including-digital-form.

[50] Umi Mokhtar. (2022) The Elusiveness of AI-based Records Classification. InterPARES Trust AI Symposium presentation, Lanzarote, Spain, October 22, 2022.

Policy measures

InterPARES Trust AI is researching the intersection of government policies and regulatory guidelines with AI technologies and trustworthy records. One of the studies on this topic is reviewing and comparing legislative and regulatory guidelines in Sweden, Finland and South Africa that inform e-government development pertaining to different AI technologies and the creation and use of records. Such study will identify key agencies and municipalities that utilise AI in e-government development, and determine recordkeeping challenges and gaps in the utilisation of AI within the realm of e-government development.[51]

Another study has reported on regulations relating to the use of AI in assisting appraisal and disposition in Latin America.[52] A policy review of ethics in AI in China is underway by researchers at Renmin University of China.[53]

National and International Cooperation

InterPARES Trust AI would not exist without national and international cooperation and multidisciplinary dialogue and exchange. One of the first items of business for the entire research team was to establish working relationships between records and archives experts and artificial intelligence scientists in academia, archival organizations, government agencies and business across international lines by ensuring mutual understanding of theoretical foundations, terminology, issues and challenges. These relationships are evidenced in each of the articles of this special issue.

Tools for use in archival processing and research/access have been and continue to be created at the UBC NLP lab for language recognition, machine translation, optical character recognition; and at the University of Macerata for image recognition. Tutorials have been and continue to be developed at UBC for archival and broader use, in natural language processing, part-of-speech tagging, named entity recognition, text classification, machine translation, and automatic speech recognition.

In addition, we are working with ISO TC 46/SC 11 on terminology, and the Canadian General Standards Board (CGSB) on the third edition of the standard "Electronic Records as Documentary Evidence," bringing AI into that standard. Several partners have created conferences or speakers' series, or invited researchers to showcase I Trust AI work, including the Sorbonne, Abu Dhabi; the National Library and Archives, Abu Dhabi; Carleton University; ENES-Morelia, Mexico. The work and influence of InterPARES has been recognized with the creation of the annual InterPARES Summer School held each June at San Benedetto del Tronto, Italy, established by the Italian Ministry of Culture, the Superintendency of Archives of Regione Marche, and the

[51] Proscovia Svärd. (2022) Investigating the Use of AI Technologies in the Realm of e-Government Development. InterPARES Trust AI Symposium presentation, Vancouver, Canada, February 20, 2023.

[52] AA01-SG05 On appraisal and disposition in Latin America. InterPARES Trust AI Report. October 14, 2023. https://interparestrustai.org/assets/public/dissemination/AA01-SG05AppraisalinLatinAmericafinaloct112023.pdf

[53] Sherry Xie. (2022) AI Ethics in China – A Policy Review. Presentation given at InterPARES Trust AI plenary meetings, Paris, France, June 29, 2022.

University of Macerata. The first Summer School, held in 2023, hosted 40 mid- and late-career professionals from 13 countries on 4 continents. Teachers were InterPARES Trust AI researchers.

Every year, three public symposia are hosted by InterPARES partners, while other initiatives to communicate the I Trust AI research include a residency in Olot, Spain in April 2023, organized by the Society of Catalan Archivists and Records Managers (AAC) in collaboration with the Ramon Llull Institute.

The breadth of I Trust AI research and the reach of dissemination can be seen in studies like the Tatuoca Magnetic Observatory, preserving the scientific memory of the Amazon. This study, with the lead researcher from Brazil and Graduate Research Assistants from UBC, has been publicly presented at workshops in Spain and Japan.

Through round table and workshop discussions, question and answer sessions at conferences and symposia and other public fora, questions, comments, and ideas from listeners and knowledge users are continuously gathered and incorporated in ongoing research and developing research design. Knowledge mobilization activities mentioned here are captured in the Dissemination section of the InterPARES Trust AI website (https://interparestrustai.org).

It has been our pleasure to compile the articles in this special issue to show just some of the work underway by the research partnership. It is our hope that they will provide inspiration and guidance to members of MoW as they continue the vital work of preserving the world's documentary heritage.