Artificial Intelligence and Archives

Shane Majors

Principles and Practices in Archives

Professor Derek Mosley

30 November 2024

Introduction

The power of Artificial Intelligence (AI) is a driving force in the advancement of many fields including that of galleries, libraries, archives and museums (GLAM). Whether or not archivists want to embrace this is not an option as it's revolutionizing the field and the future of practically every industry. Through the readings from the past two semesters, it seems cultural heritage institutions have been slow and even resistant to embracing technology to the point where most GLAMs are behind other industries and for the past decade or so have been playing a game of catch-up. AI is proving to be transformative in many aspects of the archivist's work including improving efficiency by automating daily routines, providing greater access to records and the ability to discover contexts and links that were never before imagined. We need to embrace AI and all of its powers if we are to continue to move forward; as more and more born-digital records become the norm the processing capability of AI is going to be imperative for appraisal, records maintenance and discoverability.

AI Meets Archives: The Future of Machine Learning in Cultural Heritage

In this article published by the Council on Library and Information Resources, Professor Jane Winters hints at which areas AI is currently operating and where its future uses could be employed. She states, “AI is essential for cleaning, exploring and visualizing archival and special collections...We rely on computational methods to interact with and interpret these collections.” Optical Character Recognition (OCR) is already used for making texts searchable but Handwritten Text Recognition (HTR) in combination with Natural Learning Processing (NLP) has the potential

for opening more access to manuscript collections and helping to uncover historical silences (Patton, 2024). NLP, which can identify people, places and concepts in unstructured texts, is an invaluable tool to create linked data throughout multiple collections in different institutions thus giving the researcher greater access and clearer context to those records. While the possibilities and applications of AI are almost endless, institutions do face challenges through buy-in, financial constraints, and training. Though this article is an introduction to the many facets of what AI can do and its potential, I'll delve further into the topics of exploration and visualization, text and handwriting recognition, and the challenges and ethics of AI.

AI: Exploration and Visualization

In the article, Winters mentions exploring and visualizing archival collections with AI and with the work being done by groups like the Online Computer Library Center (OCLC) in the field of linked data, visualization and exploration of collections is becoming possible. But what exactly is linked data? "Linked data is a way to organize and connect data on the web so it can be easily, automatically, and programmatically shared and used by various systems and services." (OCLC, 2024) In essence, it's about breaking down the walls created by systems and programs to make information easily discoverable and accessible. Within our current systems, if one is researching a topic, one would have to know or search for which institution is housing the records for what they are looking for. This process is not only time consuming but limits the possibilities of discovery. Imagine searching for your research topic and come across a tantalizing source, with today's systems you can only search by author, subject, etc., but imagine if that tantalizing

source was linked with another database, an event, a time period, a geographic location? Connections that would never have been found the traditional way would be revealed and expand the context of records to create a more comprehensive picture of events or people. Collections once buried in the stacks would have the possibility of being found, hidden stories coming to light.

As born-digital records are being created on a massive scale, the traditional methods of selection, appraisal, and arranging need to be reexamined to fit in the modern context of digital records and more importantly need to be automated. Machine learning is the key as it can be trained to process overly large volumes of digital records, email collections, and extract metadata (Arias-Hernandez, 2024). Automated processes are already being used on sites like Archivematica for batch processing thus making the workflow smoother and less time consuming. This is where investment and collaboration are imperative; if the field is to move forward, we need to design AI programs specifically for archival practices. In October, the National Archives announced their new framework that "charts a course for the agency that emphasizes building digital capacity, scalability, and reasonably embracing technological innovation" (National Archives News, 2024). Pilot programs have shown that by incorporating AI into their workflows, it's working to eliminate backlog and help in the efficiency of routine operations both in processing records and as an administrative business tool. As Chief Information Officer Sheena Burrell stated, "AI technology has the potential to revolutionize the way we work...By automating routine tasks and providing us with new tools to analyze and understand our data, AL can help us to be more efficient, effective, and responsive to the needs of our customers" (National Archives Nes, 2024).

AI: Text and Handwriting Recognition

Winters remarks, “HTR [Handwriting Text Recognition] has the potential to open access to manuscript collections in the ways Optical Character Recognition (OCR) did for digitizing printed text” (Patton, 2024). In my opinion, this is one of the most exciting applications of AI in the archival field. The ability, specifically of HTR, opens access on a whole different level to records we never thought imaginable. The article *Inside the AI competition that decoded an ancient Herculaneum Scroll* illustrates the unimaginable becoming reality. Discovered in 1750, the Herculaneum scrolls make up the largest intact library from antiquity surviving the elements because they were carbonized from the pyroclastic flow of the Vesuvius eruption in 79 CE. Over the centuries, any attempt made to unroll and read them ended up in a pile of ashes along with any hopes of uncovering their words.

In 2019, Brent Seals, a professor of computer science, and his team took micro-CT scans of two of the scrolls. The problem became that because the scrolls were carbonized, the carbon-based ink was almost indiscernible. A competition was opened in 2022 and scientists, computer engineers, and enthusiasts from all over the globe partook in trying to read the texts. Through progressive goals, teams were getting closer and closer when three students from different parts of the world decided to work together and build an AI model that, after nearly two thousand years, read the first words of one of those scrolls. Their model ended up revealing 2,000 characters, thus winning them the prize of $700,000. What was astonishing was the speed at which their AI model was able to identify the letters, a month, which would “usually take

papyrologists twenty years of intense study" (Weber, 2024). The applications for HTR go beyond the archaeological and into the research field; in 2022 one of the National Archives first uses of AI were to help identify names in the 1950 census. Handwritten, these documents can be difficult to read and time consuming to sift through so NARA employed AI to identify names and make the records searchable within their catalog (National Archives News, 2024). Ancestry uses similar technology on their sites as well, including Newspapers.com. One of the tenants of archival work is access and this use of HTR has ripped the door off the archive and has allowed unprecedented access to records that anyone can search. The National Archives is developing a pilot project to test AI to perform user-directed search queries. ArchieAI as it is known, is scheduled to be released this December to the public for testing and feedback (National Archives News, 2024).

An example of HTR that is available for anyone to use is Transkribus. Per their website, "Transkribus enables you to automatically recognize text easily, edit seamlessly, collaborate effortlessly, and even train your custom AI for digitizing and interpreting historical documents of any form." A downloadable app, it makes transcribing handwritten documents into readable text. An account is free, so I put it to the test: I first had it transcribe a letter from 1944 when my great grandfather was in Japan. The result was fairly accurate with but a few minor spelling errors. The second test however wasn't as successful with a letter written in Italian, it had many errors, however they do offer a "super model" version with enhanced recognition software for a fee of course. Economics aside, the leaps and bounds that AI is making changes the way information is discovered and accessed but this all comes with a cost that isn't monetary.

AI: Challenges and Ethical Concerns

Winters notes that GLAMs are slow in the adoption of AI tools which stem from financial constraints and professional development gaps. She also notes the ethical concerns around data privacy and possible biases built into AI systems and is quick to say that “it’s not humans versus machines, but humans and machines working together” (Patton, 2024) and that human skills are still essential for interpreting data and making ethical decisions. In their article, *What’s wrong with Digital Stewardship,* Blumenthal et al. surveyed organizations engaged in digital preservation and some fundamental issues were brought to light: an absence of understanding resulting in a lack of buy-in from leadership, financial constraints, and unskilled staff. Per the results, the archivists' work is governed by short-term objectives based on a grant-funded period, deliverables due in a specific timeline, and even performance indicators. “Digital preservation is a maintenance-heavy undertaking that does not... lend itself to bureaucratic performance indicators which often sway decision-making and funding priorities” (Blumenthal et al., 2020). Constantly having to go after funding to sustain their institutions initiatives is leading to burnout and frustration. By not having a thorough understanding of what digital stewardship is and what it takes and by only focusing on short-term goals and funding, leadership is failing to uphold their institutions mission with major implications regarding legal, contractual, and moral obligations (Blumenthal et al.).

Financial constraints go beyond project funding; fighting for money to maintain a digital preservation program is only part of the problem when there aren’t enough staff to carry out the institution’s mandate. This lack of manpower is an inhibiting factor, and many archivists feel the

pressure that it's all on them. In my opinion, if an institution is constantly running to find funding or technology to preserve the digital items in their collections then they need to update their mission statements and policy and not collect digital items but rather refer them to an institution that has the capability to do so. Furthermore, if an institution decides to take on a digital preservation program, then there needs to be a shift in upper management's mindset and training not just for the practitioners in the trenches but for those in control. Basic digital competencies should be a requirement and collaboration with other departments such as IT or computer science should be encouraged (Arias-Hernandez, 2024), this fostering will lead to growth and program development which is much needed in the archival community.

Ethical concerns surrounding privacy are nothing new but when coupled with born-digital records, Facebook, Instagram and any number of social media platforms, privacy seems to be a thing of the past and this is where machine learning can be utilized in the discovery of personal identification information. A perfect example of this technology already in use can be found in the application ePADD, which was developed by Stanford University in 2010 (Colavizza et al.). This software has "pioneered the application of machine learning and natural language processing to confront challenges...of screening emails for confidential, restricted, or legally protected information, preparing email for preservation, and making the resulting files...discoverable and accessible to researchers" (ePADD). The outlook for machine learning is promising, particularly around privacy concerns, but nothing is foolproof, and the human eye is needed for auditing to be sure such concerns are being met – remember it's human *and* machine.

Probably the most hidden aspect of the glory that is AI is the impact this technology has on the environment. When researching this angle, it was mind-blowing that 1) almost nothing is written on this topic in the archival field and 2) how truly detrimental it is to the environment and humans. For starters, the supply chain for the elements to create the machinery for this technology is global; mines all over the world extract silicon, copper, tungsten, quartz, gallium and tin to name a few. With the ever-growing demand for this technology, our dependance on these elements could be catastrophic with storms causing mine closures like Hurricane Helene (Osho-Williams, 2024) and historic droughts in Taiwan straining semiconductor production (Feng, 2023). Data centers are the nucleus of digital activity - the AI industry, digital preservation, your recent Facebook post, your favorite YouTube channel, your photos from your recent trip all end up there, not in the sky, as your cloud storage implies. These data centers are a massive drain on resources, particularly water and electricity. Electricity, of course, powers the enterprise while water is used to keep the servers cool. In their 2023 Annual Environmental Report, Google consumed 4.5 billion gallons of water globally in 2021 and in 2022 the number went up to 5.5 billion gallons. And that's just Google. There exists Amazon Web Services, Meta Platforms, Microsoft Azure, NTT Global Data Centers, Oracle, Equinix and CloudHQ to name a few.

In its end of life, the hardware that has made our lives easy and enjoyable, e-waste as it is called, must go somewhere and roughly 20 percent is recycled or disposed of legally, 4 percent is dumped in landfills and a staggering 76 percent makes its way illegally to poorer countries (Pendergrass et al., 2019, Favarin et al., 2023). While economically beneficial, these countries lack the environmental and safety standards of their wealthier counterparts, consequently people are exposed to arsenic, CFCs and other toxic materials. (Favarin et al., 2023) By illegally

exporting our used phones, computers, laptops, monitors, servers, processors and other electronics, we're contributing to the creation of toxic waste lands and are complicit in the harm of others and the environment.

The article *Toward Environmentally Sustainable Digital Preservation* by Pendergrass et al. offers insight into other ways archivists and institutions can take steps to becoming better stewards of the environment. One consideration starts with appraisal and that environmental costs should be included in the criteria. Another important consideration is the level of preservation that should be applied – resource-intensive, minimal preservation or low resource preservation to determine acceptable loss. (Pendergrass et al.) Not all collections are the same, therefore resource-intensive standards shouldn't be applied across the board, especially if it's a digitized version of an existing analog record. And probably the most basic and fundamental practice is that not everything should be saved, why waste time and resources on content that is truly ephemeral or duplicative. (Pendergrass et al.)

Conclusion

The future is here and we as archivists need to embrace it if we want to develop as a profession. No more are we hidden in dusty stacks but instead have the ability to be on the cutting edge of technology, and incorporating AI into our everyday routine will be a key to our success especially considering the ever-growing creation of born-digital content – the lone archivist may finally have a sidekick. The potential of machine learning is already being seen in automated tasks for efficiency in pilot programs at the National Archives and a reduced margin

of error in applications like Archivmatica and ePADD. Not all institutions are created equally and while some have the funding capabilities to undertake the endeavor right away, some can't, and the best way is to start slow. If an institution begins collecting born-digital records or plans a digitizing initiative, there are things to consider – financial capabilities, support from upper management, properly trained staff, the technology to support such an initiative, and the creation of policies concerning the environmental impact of it all. Access is a cornerstone of what we do as archivists and with the help of OTR and HTR more records are becoming available and the reading of past works that were once considered chunks of charcoal can now begin to be read again. It's an exciting time to be in the profession as more technology is created and as Professor Jane Winters stated, "It's not humans versus machines, but humans and machines working together" (Patton, 2024).

Bibliography

Arias-Hernandez, Richard. ACA 2024 Research Presentation

https://interparestrustai.org/assets/public/dissemination/AriasHernandez-ACA2024ResearchPresentation.pdf

Blumenthal, Karl-Rainer, Peggy Griesinger, Julia Y. Kim, Shira Peltzman, Vicky Steeves. 2020. "What's Wrong with Digital Stewardship: Evaluating the Organization of Digital Preservation Programs from Practitioners' Perspectives." *Journal of Contemporary Archival Studies,* 7: article 13.

Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, Julia Noordegraff. 2021 "Archives and AI: An Overview of Current Debates and Future Perspectives." *Journal on Computing and Cultural Heritage,* 15 (1): article 4.

Favarin, Serena, Giulia Berlusconi, Alberto Aziani, and Samuele Corradini. 2023. "Transnational Trafficking Networks of End-of-Life Vehicles and e-Waste." *Global Crime* 24 (3): 215–37.

Feng, Emily, "Taiwan makes tough decisions as it faces its worst drought in nearly a century." *NPR*, April 13, 2023

https://www.npr.org/2023/04/13/1169462995/taiwan-makes-tough-decisions-as-it-faces-its-worst-drought-in-nearly-a-century

Google 2023 Environmental Report

https://sustainability.google/reports/google-2023-environmental-report/

"National Archives' New Strategic Framework Emphasizes Building Capacity Through Responsible Use of Artificial Intelligence." *National Archives News.* October 17, 2024.

https://www.archives.gov/news/articles/new-strategic-framework-artificial-intelligence

OCLC. 2024. “Linked Data: The Future of Library Cataloging.” *OCLC*. https://doi.org/10.25333/71w4-vq71.

Osho-Williams, Olatunji, “Hurricane Helene Shutters 'Critical' Quartz Mines That Power the World's Electronics, Solar Panels and A.I.” *Smithsonian Magazine,* October 3, 2024

https://www.smithsonianmag.com/smart-news/hurricane-helene-shutters-critical-quartz-mines-that-power-the-worlds-electronics-solar-panels-and-ai-180985187/

Patton, Stacey. “AI Meets Archives: The Future of Machine Learning in Cultural Heritage.” *CLIR,* October 21, 2024. https://www.clir.org/2024/10/ai-meets-archives-the-future-of-machine-learning-in-cultural-heritage/

Weber, Tomas. “Inside the AI Competition That Decoded an Ancient Herculaneum Scroll.” *Scientific American.* March 19, 2024.

https://www.scientificamerican.com/article/inside-the-ai-competition-that-decoded-an-ancient-scroll-and-changed/

ePADD Project Web Site

https://www.epaddproject.org/home