**AI-Driven Digital Archiving Systems: A Study of Information Organization and Metadata Trends**

**Abstract**

This project investigates how Artificial Intelligence (AI) is transforming metadata management in digital archiving systems. It explores how AI tools like natural language processing (NLP), computer vision, optical character recognition (OCR), and large language models (LLMs) are being used by major digital repositories to enhance metadata creation and classification. By analyzing metadata from five major digital archives and comparing their use of AI tools, the study identifies common trends, challenges, and areas for improvement. The findings highlight the growing yet cautious integration of AI into archival workflows and suggest future directions for more comprehensive, context-rich metadata practices.

**Key Terms**

* Artificial Intelligence (AI)
* Metadata
* Digital Archives
* Optical Character Recognition (OCR)
* Natural Language Processing (NLP)
* Large Language Models (LLMs)

**Introduction**

**Overview:**

In today’s digital age, the amount of information being created, stored and shared has grown rapidly faster than traditional systems can handle. Archives, libraries and museums are working hard to keep up with this explosion of digital content. To do so, many have started turning to Artificial Intelligence (AI) to improve how they organize and preserve their digital collections. AI tools can help automate tasks like tagging, categorizing and describing files, making it easier to manage massive digital archives (Oyighan et al., 2024). These technologies make archiving more efficient, and they also help users find and access information more quickly and accurately.

This project focuses on understanding how AI is being used in digital archiving systems to organize information and manage metadata. Metadata, which includes details like a title, author, date and subject plays a crucial role in helping users search for and understand digital content. Traditionally, metadata has been added manually by archivists and librarians, a time consuming and labor-intensive process. Now, AI can assist with generating metadata automatically using tools like natural language processing, computer vision and large language models (Yuan, 2024 and Groppe et al., 2025).

By studying existing digital archives that use AI, this project aims to uncover trends in how metadata is created, structured and stored. It also explores how AI is reshaping archival practices and what that means for the future of information access and preservation. To do this, metadata was collected from several well-known digital archives such as Google Books, Europeana, Internet Archive and the Library of Congress, and then organized into a dataset for analysis.

**Project Questions:**

This project is guided by the following key questions:

1. How is artificial intelligence being used in digital archiving systems to organize and manage metadata?

2. What metadata fields are commonly used across different digital archives, and how are they structured?

3. What trends can be observed in AI-driven metadata generation and classification methods?

4. How do AI tools improve or challenge traditional archival practices in terms of accuracy, efficiency, and accessibility?

**Related Work**

The use of artificial intelligence in metadata management has become a growing topic in library and information science. In the article “The Role of AI in Transforming Metadata Management: Insights on Challenges, Opportunities, and Emerging Trends”, the authors explore how AI is being integrated into library systems to automate metadata tasks such as tagging, classification, and retrieval. Through a systematic literature review, the study identifies both challenges like data quality, ethical concerns, and lack of technical infrastructure and opportunities such as enhanced discoverability and automated metadata creation. The article emphasizes the importance of human oversight, noting that while AI offers efficiency, trained professionals are still essential to ensure accuracy and relevance (Oyighan et al., 2024). This work lays the foundation for understanding the balance between automation and manual curation in digital archives which is the central theme of the project.

The article “From Paper to Pixels: How Artificial Intelligence is Reshaping Archival Practices” dives deeper into how AI technologies are transforming the broader landscape of archival management. It focuses on tools such as machine learning, natural language processing (NLP) and computer vision which are increasingly used to automate repetitive tasks and enhance metadata quality. The study points out that as digital data volumes grow, AI helps archivists handle “data overload” while improving user access to materials. It also discusses long-term digital preservation strategies made possible through predictive analytics and anomaly detection. This article supports the relevance of AI in overcoming traditional archival limitations and highlights the shift from manual to intelligent workflows (Yuan, 2024). These insights align with the project’s exploration of how metadata structures vary across digital repositories and the role AI plays in those differences.

Lastly, the preprint article “Automated Archival Descriptions with Federated Intelligence of LLMs” introduces an innovative AI model that uses large language models (LLMs) to generate high-quality metadata descriptions. The authors present a federated optimization system that combines the outputs of multiple LLMs to produce more consistent and accurate archival metadata. Their experiments on real-world data demonstrate improved performance compared to traditional, single-model approaches. The paper also addresses practical concerns, such as maintaining consistency with archival standards and minimizing errors from automated systems (Groppe et al., 2025). Overall, this article offers a cutting-edge example of AI in action, showing how LLMs might be used in the future to streamline metadata creation across multiple digital platforms.

Together, these three articles provide a well-rounded understanding of how AI is shaping digital archiving today and where it might be heading. They highlight both the practical tools already in use and the experimental technologies still being developed. This background helps me build a clearer picture of how metadata is being created, managed, and improved through AI. It also supports my project’s goal of exploring current trends, understanding common challenges and suggesting ways AI can help make metadata workflows in digital archives more effective and sustainable.

**Methodology**

**Data Collection and Preprocessing:**

For this project, I collected metadata from five major digital archives: Google Books, Europeana, the Internet Archive, Open Library and the Library of Congress Digital Collections. These platforms were selected because they are well-established, publicly accessible and already integrate varying levels of AI into their metadata practices.

Each archive provides access to digital content along with associated metadata such as title, author/creator, date of creation, description, keywords/subject tags, and license type. When possible, I downloaded metadata in CSV format. In cases where direct downloads were not available, I manually gathered the information and entered it into a structured spreadsheet using Microsoft Excel.

To ensure consistency, I standardized all entries; for example, formatting all dates in YYYY format, making sure all required fields were filled in and using consistent phrasing for subject tags. The completed metadata set was saved in CSV format and uploaded to GitHub for public access and transparency.

**Data Analysis Methods:**

Once the dataset was cleaned and organized, I analyzed it to explore how different digital archives apply and structure metadata. First, I focused on five key metadata fields: Title, Author/Creator, Date of Creation, Description, and Keywords/Tags. Using Excel, I sorted and filtered entries by archive to identify the presence (in percentage) of each field across samples. To support this, I created a bar chart that visualizes the percentage of non-empty metadata fields. This helped reveal which metadata elements are consistently applied and which ones are often missing or underutilized.

Next, I examined the types of AI technologies integrated across five major digital archives: Google Books, Europeana, Internet Archive, Open Library, and the Library of Congress Digital Collections. Using a comparative content analysis approach, I reviewed publicly available documentation, official project websites and scholarly literature. I focused on identifying whether each archive uses the following AI tool categories: Natural Language Processing (NLP), Optical Character Recognition (OCR), Computer Vision, Metadata Enrichment Tools, and Large Language Models (LLMs). Based on this, I created a stacked bar chart showing which AI tools are used by each archive.

These methods helped identify both the consistency of metadata application across platforms and the evolving role of AI in enhancing archival processes.

**Result & Analysis**

The results from the metadata evaluation of five digital archives: Google Books, Europeana, Internet Archive, Open Library and the Library of Congress show clear differences in how consistently metadata fields are filled and how metadata classification is handled.

Figure 1: Bar Chart showing Most Frequently Used Metadata Fields

The bar chart of metadata completeness across five key fields reveals that Title and Keywords/Tags were present in 100% of the records analyzed. Date of Creation and Author/Creator followed closely at 95% and 90% respectively. However, the Description field was only filled 65% of the time, suggesting a notable gap in contextual metadata that could support richer discovery and access.

In addition to field completeness, the types of AI technologies integrated across five major digital archives were collected and visualized.



Figure 2: Stacked Bar Chart showing different archive's classification method

This visualization shows the presence of five types of AI tools across the archives:

* **Europeana** uses **NLP and computer vision** through initiatives like the AI4Culture Project (Europeana).
* The **Internet Archive** and **Open Library** apply **OCR** and **metadata enrichment tools** to automate part of the classification process (Internet Archive).
* The **Library of Congress** uses **OCR** and **NLP** in its digital labs and crowd-sourced transcription projects (Library of Congress).
* **Google Books** applies **OCR** to convert scanned texts into searchable data (Google).
* **Large Language Models (LLMs)** were not in active production use in any of the five archives but are being explored in research settings for future applications (Groppe et al., 2025).

This breakdown reveals that **OCR**is the most commonly used AI tool, present in all five platforms. More advanced tools like **LLMs** and **computer vision** are either still in pilot phases or limited to a few innovative platforms like Europeana.

**Discussion**

Looking at the results, it’s clear that while some metadata fields like Title and Keywords are always included, others like Description aren’t used as consistently. This missing context can affect how easily users find or understand the content. It shows that while digital archives are doing a good job at covering the basics, there’s still room to improve how they describe materials in more detail.

In terms of AI tool usage, it’s clear that basic tools like OCR are widely adopted, especially for digitizing printed materials while more sophisticated tools like NLP and computer vision are being introduced selectively. What’s also notable is the limited use of LLMs, which are still largely experimental. Overall, this reflects a cautious but growing interest in using AI to support metadata creation, with human oversight still playing a critical role to ensure accuracy, ethical standards and quality.

**Conclusion and Future Work**

**Summary of Key Insights:**

This project highlighted that AI technologies such as Optical Character Recognition (OCR) and Natural Language Processing (NLP) are actively shaping how digital archives manage and generate metadata. These tools automate essential tasks such as text extraction, keyword tagging, and basic classification, allowing archives to process large volumes of content efficiently. However, the analysis also revealed a gap between automation and contextual depth. Metadata fields like Title and Keywords are consistent across archives but more interpretive fields like Description, Subject, and Creator remain uneven filled. This underutilization suggests that while AI enhances scalability, it may still fall short in delivering detailed metadata like human input. Furthermore, while large language models (LLMs) are being explored in research, they are not yet integrated into mainstream archival workflows which indicates a significant but still emerging frontier in AI-driven archiving.

**Contribution to the Domain:**

By comparing metadata completeness and AI tool adoption across five leading digital archives, this study contributes a useful snapshot of current practices and gaps. It also offers a foundation for future metadata research, particularly for understanding how human and AI collaboration can shape richer metadata systems.

**Strengths and Limitations:**

One of the main strengths of this project lies in its structured and comparative approach. By analyzing metadata and AI integration across five major digital archives: Google Books, Europeana, Internet Archive, Open Library, and the Library of Congress, it provides a broad view of how AI is shaping metadata practices in real-world settings. The use of visualizations also helps communicate trends clearly and makes the data more accessible. Another strength is the transparency and reproducibility of the project: all collected metadata was structured and stored in a public GitHub repository, ensuring the process can be reviewed, replicated, or expanded in the future.

However, there are some limitations to acknowledge. First, the analysis is constrained by the selection of just five archives, which, while diverse, may not represent the full global range of AI adoption in digital repositories. Also, some metadata had to be collected manually due to the lack of downloadable datasets, which introduces the possibility of inconsistency or human error.

**Recommendations for Future Work:**

For future projects like this, expanding the number and diversity of archives would help provide a more depth to the perspective on AI adoption in digital archiving. One innovative idea for the future would be to explore user feedback to evaluate how AI-generated metadata affects information discovery.

**References:**

Groppe, J., Marquet, A., Walz, A., & Groppe, S. (2025). *Automated archival descriptions with federated intelligence of LLMs*. University of Lubeck & Friedrich-Ebert-Stiftung. <https://arxiv.org/pdf/2504.05711>

Oyighan, D., Ukubeyinje, E. S., David-West, B. T., & Oladokun, B. D. (2024). *The role of AI in transforming metadata management: Insights on challenges, opportunities, and emerging trends*. Asian Journal of Information Science and Technology, 14(2), 20–26.

<https://ajist.co/index.php/ajist/article/view/4277/7601>

Yuan, Y. (2024). From paper to pixels: How artificial intelligence is reshaping archival practices. Archives of Science, 8(6), 249. <https://www.omicsonline.org/open-access/from-paper-to-pixels-how-artificial-intelligence-is-reshaping-archival-practices.pdf>

**Digital Archives:**

Europeana: <https://pro.europeana.eu>

Google Books: <https://books.google.com>

Library of Congress: <https://www.loc.gov>

Open Library: <https://openlibrary.org>

Internet Archive: <https://archive.org>

**Repository Link**

The metadata, final report and the presentation can be accessed at the following link:

<https://github.com/sansbasnet/Metadata-Repository-INFO-4730/blob/main/README.md>

**Contributions**

Since I worked individually on this project, I was fully responsible for every aspect of the work. This included conducting research, collecting and preprocessing metadata from various digital archives, creating a repository, analyze and visualize the collected metadata and finally, organize all my findings into the final report and presentation.

I’ve put in my best effort to ensure the depth of research and clarity in the final analysis. I am grateful for the opportunity to work on this project, as it has significantly deepened my understanding of the role AI plays in metadata management and digital archiving. The knowledge I gained through this course has been extremely useful, and I truly appreciate how it has expanded my perspective. After receiving your review, I am eager to refine this work further and eventually publish it. Thank you!