**AI-Driven Digital Archiving Systems:**

**A Study of Information Organization & Metadata Trends**

Huzaifah Ali, Denzel Chike, Calogero Gonzales, Jonathan Portillo

INFO 4730 501 - Digital Curation and Preservation

Kate Kargozari

May 9, 2025

# Abstract

Due to artificial intelligence, also known as AI, digital archiving is changing very fast. This has transformed the creation, management, and application of metadata across different archival systems. The goal is to better organize, access, and retain information. Our project examines how artificial intelligence technologies are being implemented in libraries, museums, and large repositories. Traditional archiving methods are heavily dependent on manual metadata and subject knowledge, which can be very time consuming and inconsistent. But new artificial intelligence tools like machine learning, which predicts the future of data by using predictive models and natural language processing, also known as NLP, have been created, making it easier for metadata to be created and classification is now automated and as well scaled. It is important to know and analyze how artificial intelligence is affecting metadata practices by looking into its current developments in digital archiving systems and finding the advantages and disadvantages of these technologies. The goal is to better organize, access, and retain information, but we aim to show the major changes in how it has evolved over time. In the The Role of AI in Transforming Metadata Management: Insights on Challenges, Opportunities, and Emerging Trends and Archives, Access and Artificial Intelligence article, it mentions how artificial intelligence improves metadata generation efficiency and also introduces novel models of semantic enrichment and interoperability across digital collections (Diseiye Oyighan, Ukubeyinje, West, & Oladokun, 2024; Jaillant, 2022). This project aims for companies that have growing digital assets to use artificial intelligence to improve preservation, retrievability, and less human errors. Institutions are looking to update their digital infrastructure while upholding reliable and good archival practices will find the findings helpful.

## Key Terms

1. **Metadata -**  descriptive data attached to digital content; It includes fields like title, author, date, and tags. It helps people find and understand digital materials in archives and libraries.
2. **Artificial Intelligence (AI) -** technologies that allow machines to perform tasks that normally require human thinking. In digital preservation, AI helps automate metadata creation and clean up messy or missing data.
3. **Europeana -** a major digital archive that connects users to cultural heritage collections across Europe. It uses shared metadata standards and AI-assisted enrichment tools to improve access.
4. **Data Integration -** the process of combining information from multiple sources into one unified format. In metadata systems, it involves resolving differences between schemas or naming conventions. (Wang, 2022)
5. **Technological Solutionism -** the idea that every social or ethical problem can be solved through technology. It becomes a problem when archives over-rely on AI tools without considering fairness or user rights. (Veale, 2024)
6. **AI Regulation (AI Act) -** a European Union policy that sets rules for how AI should be used. It affects public systems like archives by requiring transparency, fairness, and protection of user rights. (Veale, 2024)

# Introduction

Over the years, artificial intelligence has progressed from theoretical use to being a very useful practical tool in digital archiving. Experimenting with machine learning models to automate thousands of digital books for organizations like the Library of Congress and big academic libraries, has led to them implementing artificial intelligence to help with their preservation and access goals. These artificial intelligence systems can handle a variety of repetitive tasks like topic modeling, keyword extraction, image tagging, and more. This allows archivists to create meaningful metadata and find hidden relationships within digital content (Tawalbeh, 2024).

The amount of digital collections from digital historical texts, to audio, and digital images has grown, requiring an urgent need for artificial technologies tools. By increasing artificial intelligence tools in digital archiving, this will help increase discoverability and better access for academic users by having consistent metadata. It also helps digital repositories be accessible by making sure the data is defined and organized. This has been proven to be helpful for global archiving projects and digital humanities projects. The best artificial intelligence tools are both machine learning and natural language processing. Machine learning uses predictive models to analyze and detect patterns across large datasets, and content can be categorized using learned behaviors. While natural language processing methods help in extracting summaries, entities, and keywords from a large amount of text. These tools are very useful when working with semi-structured and unstructured data, where standard metadata practices often fall short. Using facial recognition or object detection allows artificial intelligence to tag and categorize visual content like photos, scanned documents, illustrations, and more.

Although there are benefits but there are also challenges that artificial intelligence faces, like a lack of transparency, algorithm bias, and the need for human supervision. According to the Archives, Access and Artificial Intelligence book, some academic scholars warn against relying so much on automated systems if historical or cultural content is not correctly defined (Jaillant, 2022). So it is advised to mix both human expertise and artificial intelligence.

## Purpose

The purpose of our project is to see how artificial intelligence technologies are beneficial and are being used to automate oganizations metadata in digital archiving systems. I will look at the performance of different artificial intelligence supported repositories and compare it to manual methods to see how these tools are changing the practice of digital curation, whether they help companies get better discoverability, accessibility, and consistency for their metadata. This will help discover if the ethical use of intelligent systems is useful in information management.

Our project draws on examples from government archives, libraries, and industry leaders for artificial intelligence archiving. In the (“Artificial Intelligence Blog Series: Metadata Generation for Digital Content | Clarivate,” 2024) article, it states that artificial intelligence was more successful in automated tagging and facial recognition for photographs, while there was a decline in performance for handwritten or historical documents, causing manual review. In the Library of Congress blog by (Potter and Saccucci, 2024), they mentioned that machine learning tools are effective, but it is still very important to have a human check to make sure archival contexts are accurate.

(Tawalbeh, 2024) conducts studies on how artificial intelligence is trained and implemented can reveal and hide historical narratives. Semantic enrichment, error rate, and efficiency are more commonly recognized as one of artificial intelligence's greatest advantages. In the Archives, Access and Artificial Intelligence, which was edited by (Jaillant, 2022), it mentions that relationships between documents, generating metadata context that gives archival descriptions more depth and meaning, can be analyzed using artificial intelligence tools. A group of researchers (Diseiye Oyighan, Ejiro Sandra Ukubeyinje, West, & Bolaji David Oladokun, 2024) organized metadata using artificial intelligence and realized that it helps digital systems work better together.

Having consistent metadata helps archives, libraries, databases, and more connect and share their data. Overall, these studies show artificial intelligence's potential and limitations in digital archiving. I believe that both human help and artificial intelligence can help achieve better results, allowing professionals to focus on both analysis and digital curation, not just tagging.

## Related Works

As archives become more digital, researchers and institutions have started turning to artificial intelligence to help manage metadata more easily. However, the use of AI in digital preservation raises important questions about data quality, fairness, and how human oversight fits into the picture. Recent studies and reports have tried to answer these questions, showing both the potential and the challenges of using AI in archives.

Wang (2022) wrote a detailed report on data integration challenges, focusing mainly on health systems. While their focus was not on digital archives directly, the issues they describe apply to metadata too. They explained that one of the biggest obstacles in automated systems is inconsistency. When different platforms use different naming systems, tags, or formats for the same kind of data, it becomes difficult for AI tools to process and connect the dots. In an archive setting, this could mean that author, creator, and written by are all used differently, making it hard for AI to combine or search records correctly. Their work highlights how even small hiccups can create huge problems when trying to scale metadata systems with AI.

Another major work comes from Europeana itself. According to their official metadata guide (Europeana, 2022), the organization has worked hard to build consistent standards and controlled vocabularies. These standards are used across different libraries, museums, and archives throughout Europe. By encouraging partners to use the same metadata fields and definitions, Europeana makes it easier for AI tools to clean records, fix missing data, and connect related content. They also encourage human involvement in the process. AI tools are allowed to make suggestions or generate tags, but human experts are needed to review and approve those suggestions. This combination of automation and human validation is considered one of the best practices in modern digital preservation.

In a different study, Veale (2024) explored the legal and ethical side of AI. Their work focused on the European Union’s AI Act, a major policy that sets out rules for how AI systems can be used. The act covers things like transparency, fairness, and accountability, all of which apply to digital archives. Veale and his team warned against what they called “technological solutionism,” the idea that every problem can be fixed by technology alone. In the context of metadata, this means that relying too much on AI without considering social or ethical impacts could backfire. For example, an AI tool might accidentally tag content in a biased or incorrect way if it was trained on incomplete or one-sided data. That could cause harm to the people represented in the archive or make certain communities invisible in search results.

Veale also stressed the importance of trust. If users don’t know whether metadata was created by a person or a machine, they might start doubting the entire archive. To fix this, they recommend that public platforms disclose when AI tools were used, how those tools were trained, and how results were reviewed. This kind of transparency would not only improve fairness but also help people understand the role AI plays in preserving our history.

Together, these sources make one thing clear: AI has the power to make digital archives faster, more efficient, and easier to use. But it also comes with risks that need to be managed carefully. Projects like Europeana show that standardization and human oversight are key to making AI useful in real life. And policy work like the AI Act reminds us that archives are not just technical systems; they are social and cultural spaces too. That means people, ethics, and access all matter just as much as speed and scale.

## Methodology

To understand how artificial intelligence is affecting metadata in digital archives, our group developed a simple but focused method to collect and analyze real world data. We chose three public platforms: Europeana, Internet Archive, and Open Library. Each one represents a different approach to digital preservation. Europeana is more structured and works with cultural institutions, Internet Archive is community driven and open to all users, and Open Library blends structure with some user-generated content. We limited our data collection to items created or uploaded between 2020 and 2025. This time frame was important because it reflects the period during which AI tools became more common in digital archiving. By focusing on newer content, we hoped to get a clearer picture of how recent metadata practices compare and whether AI seemed to have made a difference.

We collected five items from each archive, totaling 15 digital entries. For each one, we recorded the following seven metadata fields:

1. **Title**: The item’s name or heading, useful for search and discovery.
2. **Author/Creator**: The person or organization responsible for the content.
3. **Date of Creation**: The year or date when the item was made or published.
4. **Description**: A short summary of the content.
5. **Tags/Keywords**: Subject labels to help organize and filter the content.
6. **Source**: Which archive the item came from.
7. **License Type**: Information about how the item may be used or shared.

First, we stored the results in a CSV spreadsheet to keep things organized. This also allowed us to run basic analytics and easily share the data within our team. We uploaded the dataset to GitHub, along with a README file that explains what our dataset is, why it matters, and how we collected it. This approach followed the best practices for data management and transparency in research.

After the metadata was collected and cleaned, we analyzed which fields were present or missing in each archive. We used Google Sheets and basic Python scripts to create bar charts, pie graphs, and a timeline. One of the most useful charts showed how often each field was completed. Not surprisingly, the title and source fields were always present, but tags and license types were often missing. This lined up with what we expected based on our background research.

Following that, we tried to estimate which archives were using AI tools and which ones were not. Based on the Europeana metadata guide (2022), we assumed that Europeana used AI assisted enhancement. The Internet Archive mostly relies on user input and manual uploads, so we marked those entries as manual. Open Library was a mix, some entries were structured, while others seemed very basic. We also tried to understand how AI might affect the quality of metadata. Since Europeana encourages shared standards and metadata consistency (Europeana, 2022), we looked for signs that its entries were more complete. In contrast, Internet Archive entries were often missing tags or license details, which we connected to what Wang (2022) said about data integration problems. Without shared standards or formatting rules, metadata can become messy and hard to automate. This supports the idea that AI tools are only as effective as the data they’re given.

To get a better idea, we thought about the influence of policy. Veale (2024) warned that AI without transparency could lead to errors or unfair results. So we looked at how often license types were listed clearly. Our analysis showed that this field was one of the most often missing, even in platforms that use AI. That made us think about the risk of relying on automation without oversight. If archives aren’t labeling content properly, AI tools might not fix the problem, they might even make it worse.

In short, our methodology focused on collecting modern metadata from real archives, comparing the presence of key fields, and identifying patterns that might reflect AI use. We used reliable tools like GitHub, Google Sheets, Excel, Google Colab, and visual charts to research our findings. As follows, we tied our work to recent research that explains why structure, human oversight, and transparency matter just as much as the technology itself.

# Results & Analysis

Once we had finished collecting metadata from our 15 selected items across Europeana, Internet Archive, and Open Library, we looked at each entry to see how complete the metadata actually was. We tracked seven fields: title, author/creator, date of creation, description, tags/keywords, source, and license type. Based on the findings, we created several charts to better visualize which fields were complete and where there were gaps. Our bar chart of field completion showed that some fields, like Title, Author/Creator, Description, and Source, were filled in across all 15 items. These are easy to find fields that most archives prioritize. However, License Type and Tags/Keywords were less consistent. License type, in particular, was often missing or marked Unknown, and one item had no tags at all. This confirmed what we found earlier, metadata standards are not always applied evenly across systems.

The pie chart showing license distribution made this even clearer. A large portion of the items either had unclear licenses or didn’t list one at all. This is important because, as explained in the Clarivate blog on metadata (2024), license fields help users know whether they’re allowed to reuse or share content. AI tools can be trained to help identify licenses, but in practice, the archives in our sample still left many entries blank. That could be due to legal concerns, system limitations, or simply oversight. We also created a timeline chart that showed the spread of item creation years from 2020 to 2025. The chart confirmed that our dataset was well distributed across time, which helped back up our goal of analyzing modern metadata practices. If most of the items had been from a single year, it might have skewed the results, but in this case, the years were spread out enough to support a fair comparison. Due to small sample size, we may not include this in our research.

Finally, based on what we know about each archive, we made a rough estimate of whether their metadata creation process was AI-driven, manual, or mixed. Europeana has been open about using AI tools to enrich metadata, so we labeled those entries as AI generated. Internet Archive, which relies heavily on user uploads, was marked as manual. Open Library was considered mixed because some entries seemed well structured, while others were missing key fields. This breakdown helped us build a pie chart showing the proportion of entries influenced by AI, manual input, or both. These visuals supported one key point: platforms that combine AI tools with clear standards and human review, like Europeana, had more complete and useful metadata. On the opposite hand, entries from the Internet Archive were more likely to be missing information. This directly supports the findings in the research from Clarivate (2024), which says that AI tools are helpful but not a complete solution. They work best when paired with structured systems and human oversight.

## Discussion

The results from our metadata analysis showed that even in the era of AI, digital archives still struggle with consistency and completeness. While some fields like title, author, and source were present in every item, others, especially tags and license information, were missing in multiple entries. This brings up questions about how artificial intelligence is really being used in digital preservation and whether it’s actually solving the problems it’s meant to fix. According to the Clarivate (2024) blog on AI generated metadata, one of the goals of using AI in archives is to automate repetitive work like tagging and classification. These tools are supposed to scan content, identify key topics, and add relevant metadata fields like tags or subjects. However, in our dataset, one item was completely missing tags, and the license field was unclear in several others. That suggests AI tools are not being applied consistently, or they are being used, but not double checked. This supports what Potter and Saccucci (2024) observed at the Library of Congress. They tested AI tools for cataloging and found that while the tools were helpful in speeding up processes, the results still needed to be reviewed by humans. AI could misinterpret context, skip important cultural information, or even assign the wrong tags if it didn't fully understand the content. Our data reflected this same idea. Even platforms that likely used AI, like Europeana, still had occasional gaps, especially in fields like license, which are harder for machines to guess without official labels.

We also noticed that platforms that rely more heavily on user submissions, like Internet Archive, had more inconsistent metadata overall. Clarivate (2024) noted that community driven platforms often lack structured workflows, which makes it hard for AI to function properly. Without standardized input formats or mandatory fields, AI models can’t learn or apply rules correctly. This could be why the Internet Archive entries had more missing information, even though the content itself was just as recent as the entries from Europeana or Open Library.

One of the most important takeaways from our discussion was that metadata quality isn’t just about technology, it’s also about trust. If a user doesn’t see a license listed, they might assume the content is copyrighted and avoid using it, even if it’s actually public domain. If AI mislabels something or adds incorrect tags, that could affect how the item appears in search results. It could also hide important voices or topics. For example, if an item about a minority community isn’t tagged properly, it may not show up when someone searches for that topic, essentially erasing it from digital visibility. In that way, metadata is more than just labels. It’s a form of representation. And AI, while helpful, still struggles with that responsibility. As Clarivate (2024) explained, AI tools are only as good as the data they’re trained on. If they’ve never seen certain languages, cultures, or formats, they might ignore or misclassify those items. That’s why archives need to be careful about how they implement automation. It can save time, but it can also introduce new kinds of mistakes.

Finally, we also have to consider the user experience. Tags and license fields are not just technical data; they affect how people interact with archives. If fields are missing or unclear, users may get frustrated or avoid using the platform. On the flip side, good metadata makes it easier to find content, cite it, and understand how it can be used. Our findings showed that archives like Europeana, which use a combination of AI and human review, had the best balance of speed and quality. But even they had room for improvement, especially with licenses. Understandably, our project shows that AI is definitely a useful tool in digital preservation, but it’s not perfect. It can help improve metadata quality, especially for large collections, but it can’t replace the role of human reviewers. Archives that want to use AI need to create clear rules, test their tools, and always include some form of human oversight. That’s the only way to make sure that digital content stays visible, accessible, and accurate, not just now, but in the future too.

# Conclusion

Throughout this project, we explored how artificial intelligence is being used to support digital preservation, with a focus on metadata creation and enrichment. We studied 15 digital items pulled from three public platforms: Europeana, Internet Archive, and Open Library. These platforms each represent different approaches, Europeana applies structured policies and some AI assistance, Internet Archive allows community uploads with minimal control, and Open Library is a mix of both. This variety gave us a balanced way to study how metadata is handled across systems and how AI plays a role in the results.

One of our most important findings was that metadata is still far from consistent, even in recent records. This is especially surprising because we limited our dataset to materials created between 2020 and 2025 a time when AI tools are widely available. Core fields like title, author, and source were present in almost every entry, but fields like tags/keywords and license type were often incomplete or missing. These gaps can make it harder for users to find, understand, or reuse digital content. While this doesn’t mean AI is failing, it does show that the tools are not being used to their full potential.

The Europeana PRO article (2023), showed that AI can perform really well when paired with human validation. In their case study, Europeana described using tools that could suggest tags or metadata enhancements automatically, with human experts stepping in to approve or reject those changes. This method produced better, more reliable metadata without removing human judgment from the process. Our own Europeana samples seemed to confirm this. Compared to Internet Archive, Europeana’s entries were more complete and easier to navigate, likely because of their AI assisted and standards-based workflows.

Additionally, the Internet Archive Blog (2023), discussed legal and ethical questions related to AI generated metadata. One reason Internet Archive often disappears license information is that assigning one incorrectly could result in legal trouble. While AI could help make guesses, the platform chooses to leave it blank unless the user specifies a license. This policy shows how complex it can be to automate metadata, especially when dealing with legal responsibilities. It also explains why our Internet Archive entries often had missing or unclear rights information.

Another important theme that came up throughout the project was the issue of fairness and representation in metadata. Metadata isn’t just about organizing content — it shapes how users discover, interpret, and interact with digital materials. If AI tools are trained on biased or incomplete datasets, they may produce tags or labels that are inaccurate or even harmful. This is especially risky when archives include content related to marginalized communities or sensitive topics. If metadata is missing, misclassified, or biased, it can make important voices harder to find — a form of digital invisibility.

Veale (2024) warned about the dangers of “technological solutionism,” which is the idea that every problem can be solved with more tech. In digital preservation, this means assuming AI will fix metadata issues automatically without thinking about who is being represented, how, and by whom. They argued that systems need to be transparent about when and how AI is used, especially in public or cultural archives. This concern was echoed in Saccucci’s (2024) interview with Library of Congress staff. Even in large, well-funded institutions, AI systems still needed human review to ensure context wasn’t lost. AI could generate basic metadata quickly, but it couldn’t understand cultural nuance, community significance, or historical complexity. Those are things that require a human touch. If archives want to maintain ethical standards, they’ll need to combine AI tools with thoughtful policies and expert validation.

We also noticed how a lack of consistent metadata affects user trust. When fields like “License Type” are marked “Unknown,” users may hesitate to download or share content, even when it’s technically free to use. This isn’t just a technical problem, it changes how people experience the archive. A simple missing label can limit access, especially for educators, researchers, and content creators who rely on clear rights information. (Internet Archive Blog, 2023) highlighted this exact issue. Their policy to avoid assigning licenses unless the user explicitly provides one might protect the platform legally, but it also shifts responsibility to the user and often results in blank fields. While understandable, this creates a situation where important content can’t be used freely because the metadata doesn’t explain how it can be used at all.

In the future, there are a few ways this project could be expanded. First, we could increase the size of our dataset. Analyzing only 15 items gave us a snapshot of how metadata is managed, but a larger sample would provide stronger patterns. If we included 50 or 100 entries per platform, we might uncover better trends, such as whether specific institutions are more consistent, or whether certain types of content like books vs. videos are more likely to have missing data. We could also explore archives that focus on non-text materials. AI tagging for images, videos, and sound files presents different challenges compared to written content. For instance, recognizing what’s in a photo or transcribing audio depends on totally different machine learning models. It would be interesting to test whether AI works better or worse on those formats and how human reviewers step in to fill the gaps. Another idea could be studying how users interact with metadata. For example, how often do people filter by license type or search by tags? Do they rely more on metadata when browsing versus when they already know what they’re looking for? Platforms like Internet Archive could offer valuable data on user behavior, which could then be used to shape better AI design and metadata strategy.

Overall, this project showed us that AI is not an easy solution for digital preservation. It’s a helpful assistant, one that can speed up repetitive tasks and uncover patterns, but it’s not a replacement for thoughtful, ethical, human-based work. Archives like Europeana are ahead of the game because they balance automation with expert input. Others, like Internet Archive, show what happens when platforms prioritize openness but leave gaps in their system. For AI to reach its full potential in this field, systems will need better training data, clearer policies, and ongoing collaboration between engineers, archivists, and communities. Metadata isn’t just information, it’s a gateway to knowledge, culture, and history. And that means we need to handle it with care, whether we’re using AI or not.

# References

Ali, A. K. (2024). *The role of AI in improving digital archiving in university libraries*. Journal of System and Management Sciences, 14(6), 455–469. <https://doi.org/10.33168/JSMS.2024.0628>

Jaillant, L. (Ed.). (2022). *Archives, access and artificial intelligence: Working with born-digital and digitized archival collections*. Bielefeld University Press. <https://doi.org/10.14361/9783839455845>

Jaillant, L., & Caputo, A. (2022). *Unlocking digital archives: Cross-disciplinary perspectives on AI and born-digital data*. AI & Society, 37, 823–835. <https://doi.org/10.1007/s00146-021-01367-x>

Kaldeli, E. (2023, November 6). *Combining AI tools with human validation to enrich cultural heritage metadata*. Europeana PRO. <https://pro.europeana.eu/post/combining-ai-tools-with-human-validation-to-enrich-cultural-heritage-metadata>

Martínez-García, M., & Hernández-Lemus, E. (2022). *Data integration challenges for machine learning in precision medicine*. Frontiers in Medicine, 8, 784455. <https://doi.org/10.3389/fmed.2021.784455>

Miernicki, M., & Ng, I. (2021). *Artificial intelligence and moral rights*. AI & Society, 36, 319–329. <https://doi.org/10.1007/s00146-020-01027-6>

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), 4277. <https://doi.org/10.70112/ajist-2024.14.2.4277>

Potter, A., & Saccucci, C. (2024, November 19). *Could artificial intelligence help catalog thousands of digital library books?* The Signal – Library of Congress. <https://blogs.loc.gov/thesignal/2024/11/could-artificial-intelligence-help-catalog-thousands-of-digital-library-books-an-interview-with-abigail-potter-and-caroline-saccucci/>

Routhier, P. M. (2023, April 28). *Internet Archive weighs in on artificial intelligence at the Copyright Office*. Internet Archive Blogs. <https://blog.archive.org/2023/04/28/internet-archive-weighs-in-on-artificial-intelligence-at-the-copyright-office/>

Saccucci, C., & Potter, A. (2025). *Assessing machine learning for cataloging at the Library of Congress*. In *New Horizons in Artificial Intelligence in Libraries*. De Gruyter Brill. <https://doi.org/10.1515/9783111336435-017>

Signal, T. (2024, March 27). *Artificial intelligence blog series: Metadata generation for digital content*. Clarivate. <https://clarivate.com/blog/artificial-intelligence-blog-series-metadata-generation-for-digital-content/>

Pham, B.-C., & Davies, S. R. (2024). *What problems is the AI act solving? Technological solutionism, fundamental rights, and trustworthiness in European AI policy*. Critical Policy Studies. <https://doi.org/10.1080/19460171.2024.2373786>

## Contributions

Denzel: Abstract, Introduction, Purpose of the Project, Visualizations, References

Huzaifah Ali: Conclusion, Metadata collection, Visualizations, References

Calogero: Discussion, Results, Visualizations, References

Jonathan; Key Terms, Methodology, Related Work, Visualizations, References