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

Huzaifah Ali, Denzel Chike, Calogero Gonzales, Jonathan Portillo

INFO 4730 - Digital Curation and Preservation

May 9, 2025

Overview

Project focus

Exploring the role of AI in Digital Archiving system

Compared three platforms: Europeana, Internet Archive, and Open Library

Main research goal

Evaluate whether AI improves metadata consistency, discoverability, and automation in real-world archives

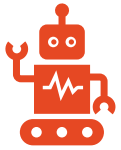
Identify AI's limitations, especially regarding ethics, bias, and oversight

Why This Matters

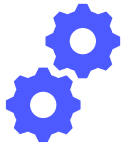
- Metadata = “Foundation of Modern Access”
  - Helps find, understand, and use digital content
- AI offers speed and scale
  - Automates repetitive tasks like:
    - Tagging
    - Sorting
    - Summarizing
  - Enhances discoverability across large digital collections
- However, AI can bring challenges
  - Quality and ethics can’t be automated
  - Incomplete training data or lack of human review leads to bias, mislabeling, or digital invisibility
- “AI tools are only as good as the data they’re trained on.” (Clarivate, 2024)

Key Terms

Metadata



Artificial Intelligence (AI)



Data Integration

Technological Solutionism

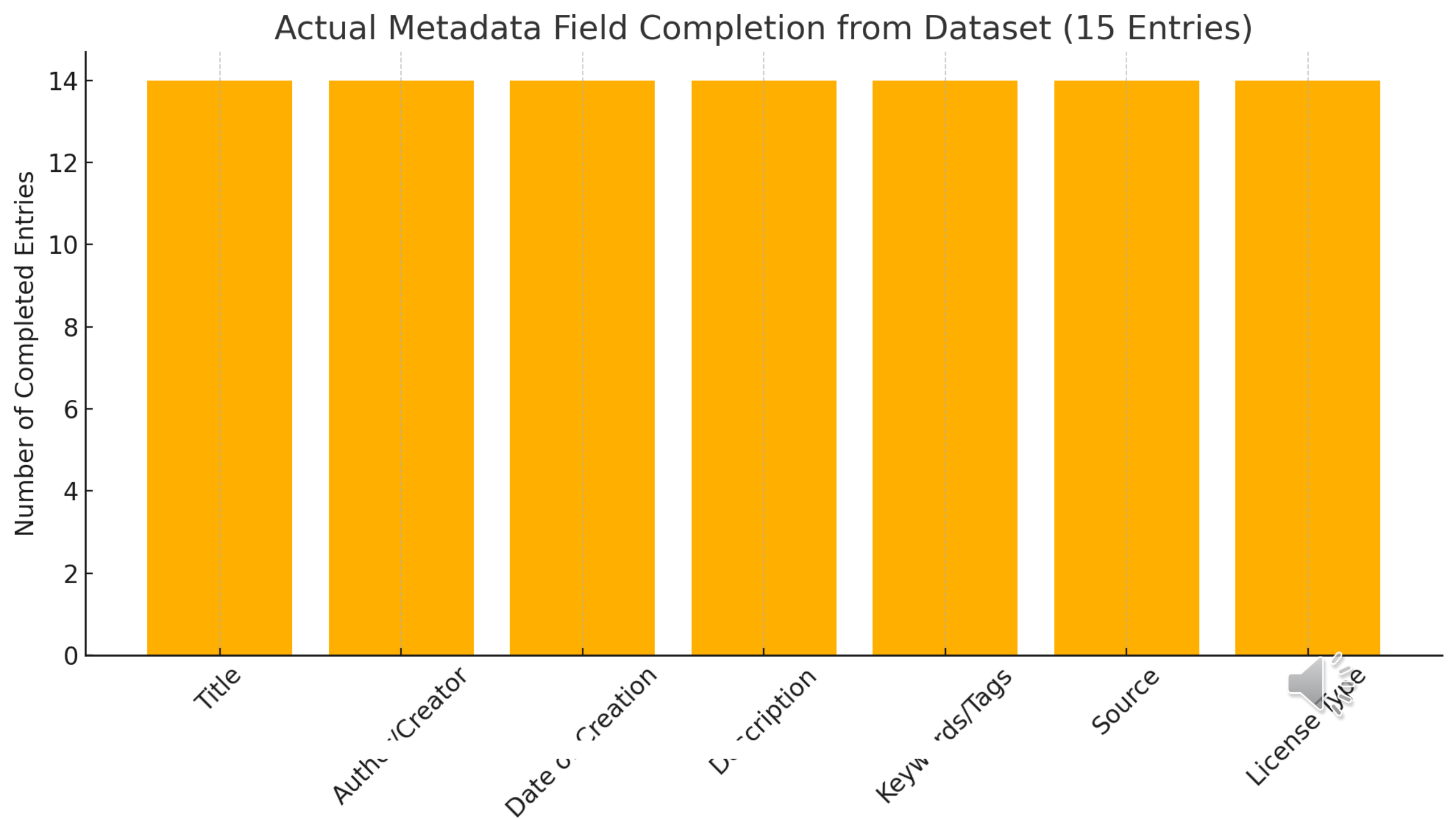
AI Regulation (AI Act)

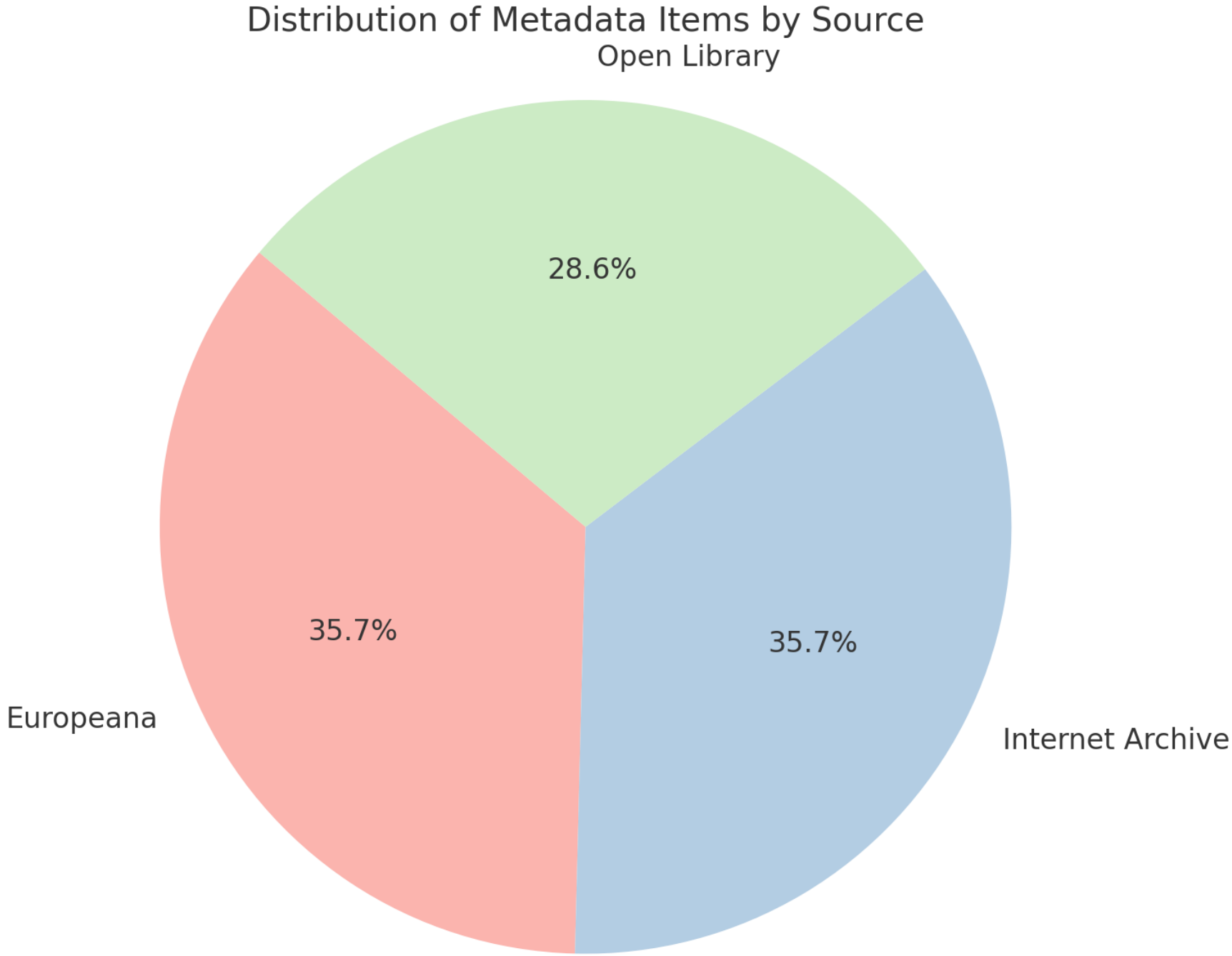
Methodology

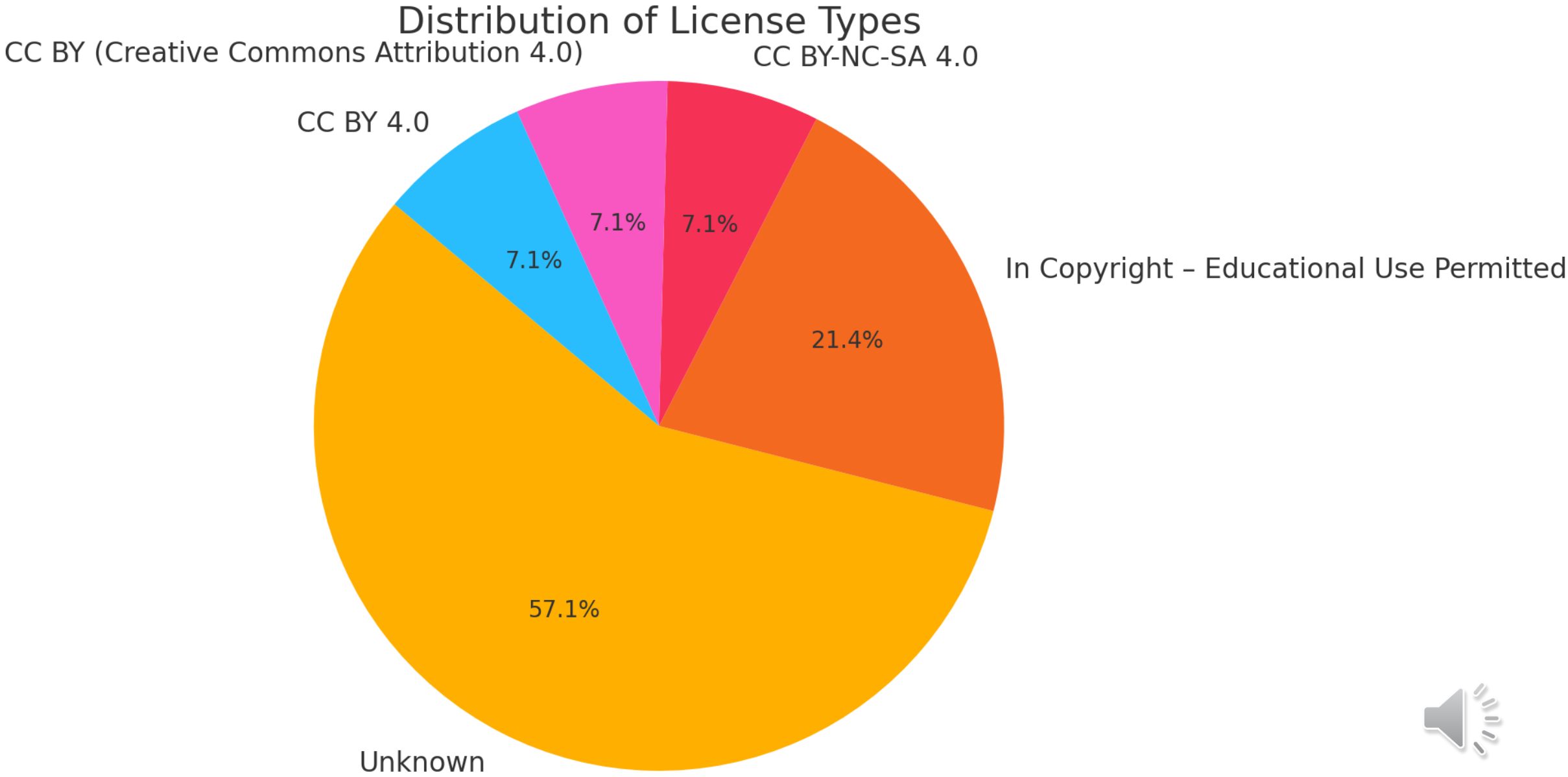
- Studied 3 public digital archives:
  - Europeana (AI-assisted)
  - Internet Archive (manual)
  - Open Library (mixed)
- Collected metadata from 15 items
  - 7 fields: Title, Author, Date, Description, Tags, Source, License
- Focused on content from 2020–2025
- Used Excel, Google Sheets (data entry); GitHub (versioning); Python (visualizations)
- Compared metadata quality and field presence to estimate automation vs. manual trends

Visualizations

- Metadata Fields Bar Chart
- Source Distribution Pie Chart
- Metadata License Type Pie Chart


Actual Metadata Field Completion from Dataset (15 Entries)
Number of Completed Entries
0
2
4
6
8
10
12
14
Title
Source


Distribution of Metadata Items by Source
Open Library
28.6%
35.7%
Europeana
35.7%
Internet Archive


Distribution of License Types
CC BY (Creative Commons Attribution 4.0)
CC BY-NC-SA 4.0
CC BY 4.0
In Copyright – Educational Use Permitted
7.1%
7.1%
7.1%
21.4%
57.1%
Unknown

Key Findings

AI improves metadata completeness only when paired with structure and human review	Europeana's hybrid system (automation + expert validation) produced the most complete entries
Inconsistent metadata in community-driven archives	Internet Archive entries lacked tags and license info due to unstructured, user-submitted data
Missing license metadata limits access and usability	Users hesitate to share or reuse content when usage rights are unclear
AI-generated tags are helpful—but not always accurate	Some items had irrelevant or missing labels, showing the need for manual checks
Good metadata = user trust, visibility, and representation	Mislabeling or omission can erase marginalized content from search and discovery

Ethical Challenges

- Algorithmic Bias in Metadata
  - AI tools may misclassify or ignore content related to marginalized communities
  - Leads to digital invisibility
- Metadata = Representation
  - Incorrect tags affect how content is found, understood, and remembered
  - Incomplete fields like “License Type” discourage reuse, even for public domain works
  - Bias in training data can distort meaning or erase histories
  - Poor metadata doesn’t just block access—it limits knowledge
- AI is fast—but can’t understand cultural nuance or historical complexity
- Veale (2024): Beware of “technological solutionism”
  - AI alone can't fix archival challenges without fairness, transparency, and human judgment

Case Insight: *Europeana*

Structured Use of AI with Human Oversight

Combines AI tools with expert validation

AI suggests tags and metadata, but humans review and approve entries

Hybrid model ensures accuracy and cultural relevance

High Metadata Quality and Completeness

Most Europeana items had:

- Complete title, author, description, source
- More consistent metadata than Internet Archive or Open Library

Metadata is cleaner due to shared controlled vocabularies and standards

Remaining Gaps: License Metadata

License Type: often missing or marked "Unknown“

Even with AI, some legal fields require manual input

Affects user trust and limits reuse of content

Conclusion

- AI is a powerful assistant, but not a replacement
  - Speeds up metadata tagging and pattern detection
  - Still needs human review to avoid errors, mislabeling, and cultural insensitivity
- Structure, transparency, and human oversight are essential
  - Without structure (e.g., Internet Archive), AI struggles with consistency and accuracy
- Metadata isn’t just data!
  - Also shapes visibility, access, and historical memory
  - Incomplete or biased metadata can erase communities or distort meaning
- AI requires transparency and fairness
- Future improvements must be collaborative

Future Direction

- Expand from 15 items to 50–100 entries per platform
- Identify stronger trends in metadata consistency, completeness, and structure
- Explore how AI handles images, audio, and video
  - Evaluate performance of tools like facial recognition and speech-to-text
- Study how users search, filter, and navigate archives
- Test AI Responsiveness to Diverse Content
  - Examine how AI treats marginalized communities or cultural materials
  - Assess bias in tag suggestions or keyword generation
- Strengthen Human + AI Collaboration

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

References cont.

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