DCPAPERS

The Impact of AI on Metadata: AI Study Group for Learning and Interdisciplinary Collaboration

Charlene Chou[†]

New York University, Division of Libraries, New York, NY, United States

Abstract

The AI Study Group was established to demystify AI and understand its impact on the metadata best practices and management. The group completed several tasks, such as testing various AI tools to assess their performance in metadata creation for diverse types of resources. Based on these discussions and tests, the group will continue learning, testing and evaluating AI tools for metadata best practices. Emphasis will be placed on ethical metadata practices and mitigating AI harms.

Keywords

metadata management, generative AI, machine learning, ChatGPT, Transformer, BERT, Claude3

1. Introduction

The community AI4LAM (Artificial Intelligence for Libraries, Archives & Museums) started its annual conference, Fantastic Futures, in 2018.[1] The author began testing AI/machine learning models after attending the 2019 conference. ChatGPT was launched on November 30, 2022. Since then, we have been able to communicate with machines using natural languages. It was the first day of the generative AI era. In December 2023, the author established the AI Study Group in the Department of Knowledge Access, which is responsible for cataloging and metadata services. The goal was to demystify AI and to deepen understanding of generative AI tools, in the context of metadata best practices and management. Furthermore, the author participated in the PCC Task Group on Strategic Planning for AI and Machine Learning and conducted an environmental scan on the impact of AI and machine learning on library metadata operations. The final report provided evidence about potential directions.[2]

2. Goals & research questions

The primary goal of this study group is to understand the impact of AI on metadata best practice and management through reading, testing and training. There are two overarching research questions to address, which underpin group discussions and hands-on tests. First, can AI tools assist in metadata operations such as metadata creation and quality control? For example, can we automate metadata remediation, such as geospatial metadata, for system migration? Can we use AI tools to assist in research and create certain metadata? Secondly, can AI/ML tools enhance the discovery of library resources,

*Corresponding author.

[†] 0000-0003-4736-7662 (C. Chou)

https://doi.org/10.23106/dcmi.952409098

including for multilingual or non-Latin script resources? For instance, can we enhance subject indexing by using machine learning algorithms to identify text similarities?

3. Completed tasks

The author compiled and keeps updating a bibliography of AI resources for learning purpose.[3] The group completed several tasks in three categories: reading and discussing journal articles, testing AI tools for library metadata and learning prompt engineering. First, we discussed two journal articles to evaluate whether the MARC records were correctly created by ChatGPT and to understand if BERT/transformer models can assist in automatic subject indexing for digital resources.[4] Secondly, a colleague successfully used ChatGPT to instruct how to use Python scripts to remediate metadata format and structure for 4000 JSON files into structured-format files required for the system upgrade of our Spatial Data Repository.

Regarding MARC data, we experimented with using English abstracts, Arabic abstracts, or rare book dealer descriptions to create MARC records or LC subject headings. Our goal was to determine what was effective and what was not. For instance, ChatGPT did not perform consistently for LC subject headings. On the other hand, ChatGPT was able to suggest a correct LC classification numbers for corresponding LC subject headings and we achieved better results with different queries. We also used different AI tools, such as Claude3,[5] to translate and transliterate non-Latin scripts. Lastly, the more we tested, the more we realized the importance of prompt engineering. In June, a training session on prompt engineering was conducted, and a participant requested a hands-on training session in the fall.

4. Work in progress

We will continue to maintain this group learning space to support interdisciplinary collaboration between metadata librarians, subject librarians and technologists. Our focus will be on learning and testing various AI tools to strategize metadata creation and quality control. However, our group is also committed to address ethical concerns such as bias. Therefore, we aim to test Transformer models for creating training datasets, which may help mitigate AI harms such as hallucination, bias and unethical data. Ultimately, we can leverage AI tools to assist in our metadata management and best practices, while emphasizing ethical metadata.

5. Conclusion

At the inception of this study group, our group members had varying levels of AI knowledge and experiences. After a few meetings, more colleagues were willing to participate in testing, discussions and suggesting ideas for future meetings. This is the most gratifying outcome, as it achieved the goal of creating a safe learning space for change management and demystifying AI. AI, machine learning and big data are inevitable.[6] Metadata librarians need to reimagine our roles and learn how to collaborate with diverse stakeholders, such as data scientists. We have a lot to offer due to our domain knowledge, including rich vocabularies, ontologies and metadata standards. In response to NYU strategic pathways for interdisciplinary collaboration,[7] we will proactively participate in ethical AI projects creating quality datasets and metadata and contribute our perspectives on using generative AI tools for ethical metadata management.

https://doi.org/10.23106/dcmi.952409098

References

[1] AI4LAM (Artificial Intelligence for Libraries, Archives & Museums. URL: https://sites.google.com/view/ai4lam

[2] PCC Task Group on AI and Machine Learning for Cataloging and Metadata: Final Report. URL: https://www.loc.gov/aba/pcc/taskgroup/TG-Strategic-Planning-AI-final-report.pdf

[3] AI Resources. URL: https://docs.google.com/document/d/1FeNjxXq7tw1GTgmZK-aDeG5HYwh-FZeRGTZqmDc22Ew/edit?usp=sharing

[4] Chou, Charlene, & Chu, Tony. "An Analysis of BERT (NLP) for Assisted Subject Indexing for Project Gutenberg." *Cataloging & Classification Quarterly 60.8* (2022), 807–835. Doi:10.1080/01639374.2022.2138666

[5] Claude 3. URL: https://www.anthropic.com/claude

[6] Tang, Kwok-leong, Crafting Effective Inquiries in ChatGPT and More!: A Hands-on Workshop for East Asian Librarians, 2024. URL: https://fccsdci.notion.site/Crafting-Effective-Inquiries-in-ChatGPT-and-More-A-Hands-on-Workshop-for-East-Asian-Librarians-bbba13be1c044ab8bf0ca64511712786#1abacdd1ec4842ae9d28cb5771541480

[7] New York University, Office of the President, Strategic Pathways. URL: https://www.nyu.edu/about/leadership-university-administration/office-of-the-president/strategic-pathways.html

https://doi.org/10.23106/dcmi.952409098