Using AI to Facilitate Discoverability and Curation of the ASU Library Repository Collections

Timothy Provenzano[1,*,†], Rachel Fernandez[1,†], Chad Deets[1,†] and Deirdre Kirmis[1,†]

[1]*Arizona State University Library*

Abstract

In summer 2024, as part of Arizona State University's collaboration with OpenAI, the ASU Library launched a pilot project using the AI tool ChatGPT. This project aims to enhance the discoverability and curation of digital collections within the library's repository ecosystem. The use of AI in libraries is gaining attention, with many institutions exploring AI for generating descriptive metadata. ASU Library's extensive repository platforms, including an institutional repository, data repository, and a digital collections platform, hold approximately over 10,000 objects, with numbers expected to grow. The library lacks a dedicated position for creating metadata, with the responsibility distributed among various units already tasked with other duties. This project aims to determine whether ChatGPT can effectively generate accurate metadata that meets best practices. The library will use an existing archival collection of government documents, which already has human-created metadata, as a benchmark, in comparing the generated metadata for the fields Title, Description, and Keywords. By comparing ChatGPT-generated metadata to the existing metadata, the library will assess the relevance of AI outputs and the level of oversight required. If the AI-generated metadata shows minimal variance from the human-created metadata, the workflow could expand to other collections and reduce the backlog of unpublished archival collections that require descriptive metadata.

Keywords

artificial intelligence, AI, institutional repository, project report

1. Introduction

In the summer of 2024, ASU Library is undertaking a pilot project using the Artificial Intelligence (AI) tool ChatGPT. The purpose of this pilot is to investigate the suitability of ChatGPT to facilitate discoverability and curation of digital collections within the ASU Library repository ecosystem. A considerable challenge at ASU Library is the large amount of digital material that needs to be described, at least minimally. Our resources do not currently permit a dedicated position to create metadata at the scale required. Therefore, ASU Library took part in the Arizona State University's AI Innovation Challenge, a university-wide initiative in collaboration with OpenAI to grant free ChatGPT Enterprise licenses to ASU faculty, staff, and students with ideas to positively impact the future of education.

DCMI-2024 International Conference on Dublin Core and Metadata Applications

*Corresponding author.

†These authors contributed equally.

timothy.provenzano@asu.edu (T. Provenzano); rachel.fernandez.1@asu.edu (R. Fernandez); chad.deets@asu.edu (C. Deets); dckirmis@asu.edu (D. Kirmis)

0009-0003-0246-4175 (T. Provenzano); 0000-0002-7697-4149 (R. Fernandez); 0009-0009-8457-6253 (C. Deets); 0009-0007-8262-0463 (D. Kirmis)

Our project poses the question of whether ChatGPT can positively impact our ability to generate accurate metadata that aligns with relevant best practices. For this project, we are using an existing archival collection of government documents, for which human-created metadata already exists. We will be comparing the AI generated metadata to manually-created metadata to test the relevancy of outputs and determine the minimum level of oversight. Through participation in this Innovation Challenge, we are hoping to determine if the metadata generated through AI is within a tolerable level of variance from the existing metadata and thus can be expanded to other collections.

2. Background

The use of artificial intelligence in libraries and archives is a topic of increasing attention, as the technology evolves and becomes at least a *prima facie* option for different use cases [1, 2, 3, 4]. The use of AI to generate descriptive metadata in particular has already been attempted by many institutions, and their example provided an impetus for our library to attempt this solution [5, 6, 7, 8]. Our challenges are doubtless familiar to many in academic libraries: our ecosystem of repository platforms, encompassing an Institutional Repository (IR), Research Data Repository (RDR), and a digital collections platform.

The amount of holdings across our repositories is quite large—approximately 10,000 objects in the IR and digital collections alone, as of May 2024. A steady influx of new items exists at present, and we expect this to only grow over time. At the same time, the Library does not have a dedicated position for creating object-level metadata for digital objects, and the responsibility is currently shared between different units that are already charged with other tasks.

In early 2024, Arizona State University became the first university to partner with OpenAI which has allowed for the use of ChatGPT enterprise licenses [9]. Soon after this partnership was announced, ASU created the AI Innovation Challenge and began accepting proposals from faculty and staff who wished to use a license for ChatGPT Enterprise at no cost, to investigate how this tool could assist with the learning and engagement goals of the university at large. These projects encompass a wide variety of formats and circumstances, and the university has been at the forefront of experimenting with AI tools to leverage solutions at scale.

Accepted as part of the second round of projects for the challenge, we hope to use this time and access to the enterprise license to examine the ways we can incorporate AI within our workflows at the ASU Library. If we are were to find some level of success with our project, we are confident that the Library would find institutional support to operationalize our workflow.

3. Project Outline

We will be using an existing collection of government publications as part of our our training corpus. The State and Local Arizona Documents (SALAD) collection contains documents published by the State of Arizona, county and municipal governments, affiliated councils and agencies [10]. ASU is a primary collector of state publications and makes a concerted effort to acquire and catalog most materials published by state and local governmental agencies. This collection has extant descriptive metadata, which we will compare to the outputs of ChatGPT.

Our project will involve sampling a limited set of documents from the collection and prompting the tool to generate output for the metadata fields: Title, Description, Keywords. When these entries are generated, they will be analyzed to measure both their commonality with the existing metadata, as well as their fidelity to the source material. We anticipate the evaluation mechanism to take shape during the course of the investigation, in the form of a question rubric to guide those who are reviewing the metadata for accuracy, as well as investigating the possibility of more automated or quantitative analysis. An important consideration for the project is how much oversight will the tool need, either in the pre-ingest portion of the workflow, as well as the need to do manual oversight of the output, prior to publication. Given that the initial prompt for the project was a lack of available staff or librarian positions, any requirement of extensive intervention or remediation would be less likely to result in adaptation of the tool.

References

[1] L. Jaillant, K. Aske, Are users of digital archives ready for the AI era? obstacles to the application of computational research methods and new opportunities, Journal on Computing and Cultural Heritage 16 (2024) 1–16. doi:10.1145/3631125.

[2] S. Mannheimer, D. Rossmann, J. Clark, Y. Shorish, N. Bond, H. S. Kettler, B. Sheehey, S. W. H. Young, Introduction to the special issue: Responsible AI in libraries and archives, Journal of eScience Librarianship 13 (2024). doi:10.7191/jeslib.860.

[3] G. Colavizza, T. Blanke, C. Jeurgens, J. Noordegraaf, Archives and AI: An overview of current debates and future perspectives, Journal on Computing and Cultural Heritage 15 (2022) 1–15. doi:10.1145/3479010.

[4] M. Hosseini, K. Holmes, The evolution of library workplaces and workflows via generative AI, College & Research Libraries 84 (2023). doi:10.5860/crl.84.6.836.

[5] C. Haffenden, E. Fano, M. Malmsten, L. Börjeson, Making and using AI in the library: Creating a BERT model at the national library of sweden, College & Research Libraries 84 (2023). doi:10.5860/crl.84.1.30.

[6] C. McIrvin, C. Miller, D. Smith-Glaviana, W. N. Ng, Automatic expansion of metadata standards for historic costume collections, Journal of eScience Librarianship 13 (2024). doi:10.7191/jeslib.845.

[7] E. Prokop, X. Han, V. Papyan, D. L. Donoho, C. R. Johnson, AI and the digitized photoarchive: Promoting access and discoverability, Art Documentation: Journal of the Art Libraries Society of North America 40 (2021) 1–20. doi:10.1086/714604.

[8] I. Reiche, The viability of using an open source locally hosted AI for creating metadata in digital image collections, The Code4Lib Journal (2023). URL: https://journal.code4lib.org/articles/17186.

[9] J. Boehm, R. Heath, OpenAI launches first university partnership with Arizona State, Axios (2024). URL: https://www.axios.com/local/phoenix/2024/01/18/openai-asu-university-partnership-school-arizona-chatgpt.

[10] ASU Digital Repository, Arizona State and Local Government Documents Collection (SALAD), 2021. URL: http://hdl.handle.net/2286/R.C.16, accessed: 2024-05-30.