Introduction

Northwestern University Libraries (NUL) proposes a 2-year National Leadership Grant implementation project called "Starting the Conversation: Enriching Collection Access and Use with Generative AI." This project will implement a semantic chat-based discovery tool and automated metadata creation tool using generative AI. The tools will be implemented in NUL's repository suite, which includes an in-house developed open-source digitized collections front-end application and preservation-first digital asset management system. This implementation ensures that the project is not a prototype, but a well-tested production product ("Digital Collections") with hundreds of thousands of assets and a global user base of, on average, 10,000 users and 12,000 sessions a month. Another unique feature of this project leverages generative AI tools to augment the item-level descriptions of digitized unique and distinctive works en masse. In addition, toolkits comprised of research methods, examples, and open-source software will be released for other library professionals and researchers to use and implement in their own context. This project addresses the need for academic research libraries to retain ownership and demonstrate expertise in the development of AI tools. This project directly supports NLG goal/objectives 1.1, 3.1, 3.2, and 3.3 by exponentially increasing access to and providing interpretation of information from primary resources with minimal human intervention. NUL is respectfully requesting $458,412 IMLS for this implementation project and NUL will cost share $458,615.

Project Justification

Over the last six months a seismic shift has occurred in computing. Specifically, generative AI tools—now commercially available and in use throughout higher education—requires all industries to rethink assumptions about workflows and end user expectations while digging into ethical questions. For libraries, there is "immense potential for it to increase access to knowledge in fundamental ways, for example through improved search and recommendation, through description of digital materials at scale, through transcription, and through automated translation...and there is a recurrent fear that AI may in some way replace human librarians' work." [1]

It is in this landscape that Northwestern University Libraries proposes to extend our proof of concept and build a chat-based interaction with Digital Collections, which leverages a vector-based semantic search and a large language model's (LLM) conversational ability to surface relevant information in a collection based on the meaning of the user's input. This is vastly different and far more powerful than traditional "search" utilities that rely on a keyword index to match results and a filtering or faceting system to refine results. The proposed conversational approach promises nuanced results based on context and semantic meaning of the queries. Grounding the queries of a state-of-the-art LLM to a semantic model of our collection drastically changes the concept of traditional library resource discovery and offers end users conversational context about the resources and insights into why those works surfaced.

In addition, this project will enable significantly streamlined metadata creation by using generative AI tools in new ways to augment the item-level descriptions of unique and distinctive works. These automated descriptions will allow end users to understand a corpus in new ways. The chat-based summaries and meaningful conversations with substantial amounts of machine generated metadata and transcriptions from text, images, audio, and video will make digitized collections that were once only minimally described more accessible, which advances digital inclusion (NLG goal 3.1). Both aspects of this project provide an innovative solution to managing digital collections (NLG Goal 3.2) and meets the changing user expectations of discovering and finding information via conversations in a chat (NLG Goal 3.3). Lastly, all this work will be packaged in a variety of formats: as a toolkit other libraries can implement on their own collections, as a resource for students, and as a workshop for the digital humanities (NLG Goal 1.1). These packages will also address ethical questions around this work and outline the risks involved. This project leverages the staff expertise already developed at NUL to increase access to knowledge in new ways and remove the fear around AI in libraries.

Project Work Plan

Northwestern manages its projects in the open and uses best practices from Scrum. All work is visible to the team, management, and stakeholders via GitHub issues and open presentations (demos) every 8-week cycle. Work is broken out into high-level goals and the team working on the project focuses on iterative deliverables. In addition to the project director, the project team will consist of cross-functional developers, a metadata subject matter expert, communication professionals, user experience professional, and project manager. Some of Focus 1 and 2 will occur concurrently for the

---

[1] Cox, A. M., & Mazumdar, S. (2022). Defining artificial intelligence for librarians. *Journal of Librarianship and Information Science*, *0 (0)*. https://doi.org/10.1177/09610006221142029

project term, since refining and maintaining are built into our development processes. Focus 3 will occur towards the end of year 1 and ramp up in year 2.

Focus 1: Semantic Search (Months 1-8, 14-24). Goals: Extend our proof-of-concept chat-based semantic discovery tool and move it into production. Develop a suite of reusable components as a toolkit.

Activities include: 1) Experiment with several commercial and opensource LLMs and select (Llama2, Claude, ChatGPT); **2)** Populating vector databases with Digital Collections material; **3)** Extend Digital Collections search feature to incorporate chat-based discovery; **4)** Analyze impact on collection usage; **5)** Develop tests that validate LLM responses for accuracy and validity; **6)** Use ML/AI metadata to augment semantic model.

Focus 2: Metadata Generation and Validation (Months 9 –14, 14-24). Goals: Develop software that augments human metadata workflows with LLMs (ChatGPT) to describe objects and validate against standard controlled vocabularies, including Library of Congress (LoC) and Getty. Develop a toolkit for other libraries.

Activities include: 1) Use best-in-class commercial machine learning tools to transcribe text, video, audio, and tag content; **2)** Populating vector databases with subject headings including LoC and Getty; **3)** Develop framework (create rules and constraints) that uses minimal human effort to generate descriptions and controlled vocabularies; **4)** Develop methods of validation and reinforcement against domain-specific information and knowledge; **5)** Test workflows with subject matter experts, refine methods based on feedback; **6)** Incorporate generative metadata creation into digitization workflows and monitor effectiveness, efficiency improvements, and accuracy; **7)** Develop best practices that describe methods, challenges, and risks of generative AI for item-level description. Although we did not include a diversity plan, we fully expect this set of best practices to include information about biases in these tools. Furthermore, we hope to test utilizing the tool to analyze and audit collections for outdated terms in order to efficiently remediate them.

Focus 3: Promotion of Focus 1 and 2 outputs (Months 8-24). Goals: Communicate about toolkits to other libraries. Develop a series of open education resources to enhance the Digital Humanities community understanding of large language models and promote outputs from the project.

Activities: 1) Create a faculty and professional advisory board to assist in creation of the following: an open education resource that utilizes code notebooks (Jupyter) to give faculty and students hands-on experience writing simple programs using NUL digital collections as an open dataset; and a workshop focused on generative AI in the humanities. **2)** Hire a technical communications consultant to work with NUL's Marketing and Communication team to define vehicles for outreach and develop technical and non-technical content, training materials, and articles specific to those vehicles. These will include but not be limited to: conference presentations, Northwestern publications, external technical publications (AWS blog), and community publications. NUL knows from experience with grant projects that having a strategy and time for marketing is critical for the project's success.

Project Results

As a result of this project, NUL will be running the discovery tool in production, as well as using automated metadata workflows in its expanding digital collections operation. However, the most important parts of this project are the semantic discovery and automated metadata openly accessible toolkits for other libraries to adapt to their own environments and the hands-on training materials for scholars. NUL expects to provide education and its expertise developed in the project to teach other libraries about these new technologies. The toolkits and open education resources, their promotion, and workshop address the need for libraries to retain ownership of and demonstrate expertise in the development of AI tools. Critically, this project will showcase generative AI's potential to create inclusive descriptive methodologies by exposing historic bias. Lastly, by exponentially expanding access to and providing interpretation of information from primary resources, this project increases access to knowledge and removes the fear around AI tools, allowing libraries to adapt and meet rapidly changing expectations.

Budget Summary

The total cost of the project is $916,977, of which $458,362 is requested from the IMLS, and $458,615 will be cost shared by NUL. For the requested IMLS portion, $197,671 of the requested funds will be used for salaries and wages; $54,360 will be used for fringe benefits; $23,000 for travel; $42,000 for supplies and materials; $20,000 for other costs; and $121,331 for indirect costs at a rate of 36%. For the cost share, $240,954 will be used for salaries and wages; $66,262 for fringe benefits; $30,000 for supplies; and $121,398 for indirect costs at a rate of 36%.