

1. Project Justification

Introduction

Northwestern University Libraries (NUL) proposes a 2-year National Leadership Grant implementation project called “Shaping the Conversation: Enriching Collection Access and Use with Generative AI.” This project leverages the power of generative AI to revolutionize the way libraries provide access to knowledge in two ways: 1) the project will build an installable open-source discovery product (OSDP) that allows users to chat with collections; 2) the project will test and validate a metadata toolkit to augment and suggest metadata for digitized collections. This project teaches libraries to utilize and assess generative AI technologies “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

Specifically, the OSDP will be designed as a starting point for institutions seeking to implement AI-driven semantic search for end users alongside or in lieu of existing search tooling. This type of search tool provides nuanced results based on semantic meaning of the queries, essentially analyzing the input for its meaning and intent, and then offers context for returned results. Additionally, NUL will build a metadata creation and subject heading assignment co-pilot that utilizes generative AI and related technologies. The published open-source toolkit will showcase a generative metadata pipeline in a digital collections environment, giving library professionals the skills and knowledge necessary to implement a customized version of the pipeline in their own environment. NUL respectfully requests \$439,597 from the IMLS for this implementation project and NUL will cost share \$460,820.

Large library vendors have started projects that will integrate generative AI discovery and metadata tools into their products as shown in the below footnote, and the library community, as information professionals, has an ethical responsibility to understand how these new products work.² This project intends to pull back the curtain on these new AI tools by creating the above-mentioned openly available deliverables and by offering a workshop, with detailed documentation and learning opportunities for adoption. The project directly supports NLG goal 3 and objectives 3.2 and 3.3 by increasing access to collections and meeting new user expectations in the innovative ways generative AI provides. Additionally, by ensuring libraries understand how to deploy generative AI to enhance equitable outcomes through the project’s deliverables (OSDP and metadata toolkit), the project will also meet NLG goal 1 of building the workforce’s expertise around these new generative AI tools.

The Need: Deliverable One, OSDP

Market research in changing user expectations around information seeking behavior is noted in a recent consumer survey of more than 15,000 internet users ages 16-64 and shows that “1 in 5 consumers use ChatGPT or another AI tool to find information online may not seem high, but in one year it’s already outpaced technologies that’ve been around for a decade.”³ This new information seeking behavior creates vastly different expectations for interaction than traditional library search systems, which rely on a keyword index to match results and a filtering or faceting system to refine results. The proposed OSDP seeks to meet this new user demand by applying the latest generative AI tools to library discovery solutions. Modeled after soon to be released chat-based search

¹ <https://doi.org/10.1177/09610006221142029>

² <https://exlibrisgroup.com/blog/artificial-intelligence-blog-series-prioritizing-conversational-discovery-at-ex-libris/>; <https://www.jstor.org/generative-ai-faq>; <https://law.lexisnexis.com/Lexis-Plus-AI-Launch-PPC-Google-Branded> ; <https://www.oclc.org/en/news/announcements/2023/leveraging-machine-learning-for-worldcat-de-duplication.html>; <https://proquest.syndetics.com/news/2023/06/13/introducing-talpa-search/>

³ <https://blog.gwi.com/trends/chatgpt/>

for NUL’s Digital Collections, the OSDP will leverage a vector-based semantic search and a large language model’s (LLM) conversational ability to surface relevant information based on a user’s plain-language input.⁴ The power of generative AI assisted search is best shown by example from NUL Digital Collections. For example, a user working on a paper directly asking a question in natural language allows for nuance in question forming and response not possible when trying to encode a question into keyword, operators, and filters. A researcher might enter directly into chat: “What posters in the collection depict the importance of food during war time? Are there any common themes in the posters?” Another example of a chat-based query is: “I’m writing a paper about apartheid, what kind of documents might be helpful?” (Figures 1 & 2). The main elements in NUL’s current application displays results along with contextual information based on the end user question, such as pulling out themes of the result set, in addition to thumbnails from the digitized works in the result list. This is done by grounding AI responses to the results from a semantic vector search of a collection --a method called Retrieval Augmented Generation (RAG). This approach ensures that hallucinations found in general-purpose chat-tools are mitigated.⁵

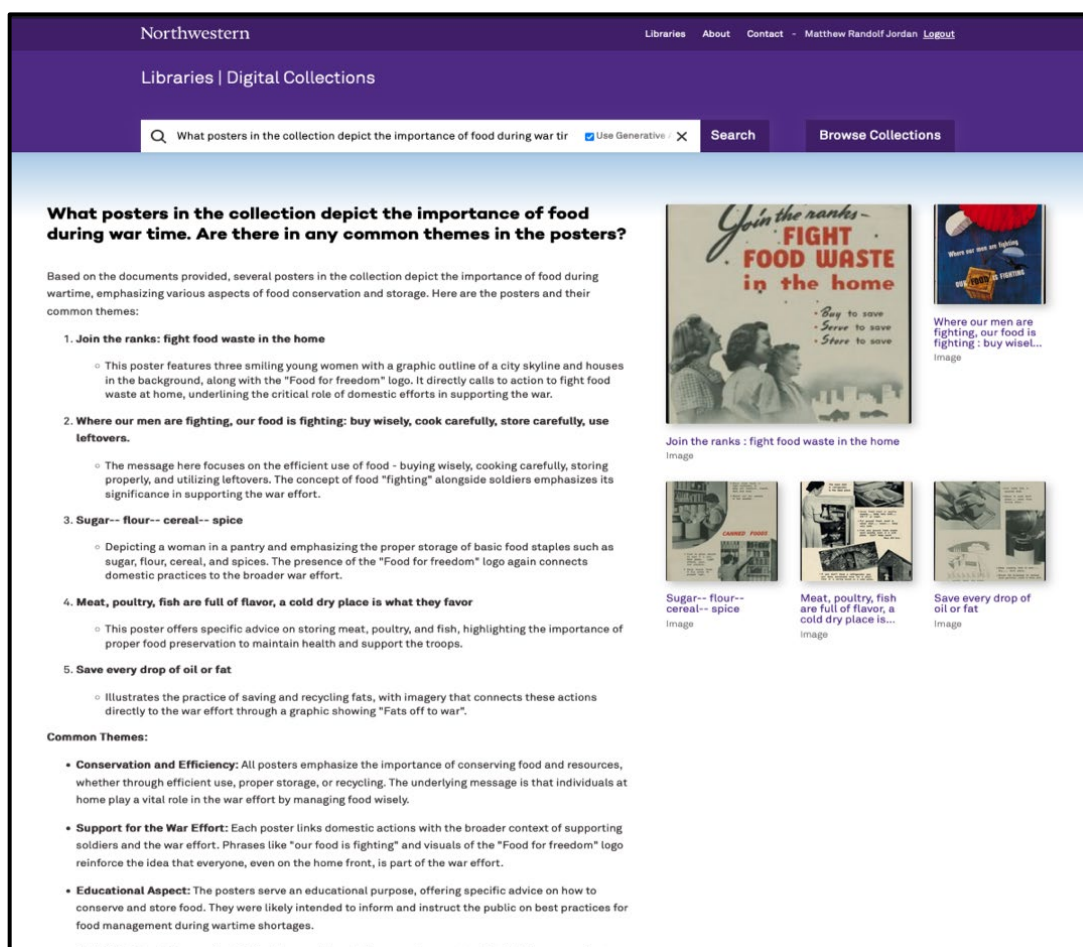


Figure 1. This is an example from NUL’s current chat application. Responses are unique to the user’s question. This response includes 3 key elements: 1) a listing of the top returned results as well as a short description of each result and how it may answer the user’s question; 2) the LLM analyzes metadata to answer the user’s question about common themes; 3) the right side shows thumbnails of the returned results from the semantic search.

⁴ <https://dc.library.northwestern.edu/>

⁵ <https://arxiv.org/abs/2312.10997>

Northwestern Libraries | Digital Collections

Libraries About Contact - Matthew Randolph Jordan Logout

Q I'm writing a paper about apartheid, what kind of documents might be helpful Use Generative X Search Browse Collections

I'm writing a paper about apartheid, what kind of documents might be helpful?

For a paper about apartheid, the provided documents offer a rich variety of perspectives, focusing on the international and grassroots movements against the apartheid regime in South Africa, the use of economic sanctions as a means to pressure the South African government, and the broader cultural and political ramifications of apartheid. Here's how each document could be relevant:

- Apartheid's friends and allies:**
 - Content:** This document features a poster that includes photographs of anti-apartheid protests and weapons, along with statistics and articles on international arms, oil, investment, and economic support for apartheid.
 - Relevance:** It could provide insight into the international dimension of apartheid, highlighting how various countries and entities either supported or opposed the regime through economic means. The document could serve as a starting point for discussing the global anti-apartheid movement and the complexities of international politics surrounding the issue.
- Racism is legal in South Africa:**
 - Relevance:** Although the description is sparse, the title suggests an exploration of how apartheid influenced or was represented in South African literature. This document might offer perspectives on the cultural impact of apartheid, examining how racism and segregation were depicted or challenged through literary works.
- Fight apartheid: protect our most precious resource:**
 - Content:** Features figures of black children, linked together as if in a paper chain, with rips and holes.
 - Relevance:** This poster could be used to discuss the impact of apartheid on children and the future generation. It symbolizes the fragility and interconnectedness of lives under the oppressive regime, offering a poignant visual representation of the social and human cost of apartheid.
- Save the Sharpeville 6, no to apartheid executions:**
 - Content:** Graphic design of two hands breaking apart the rope of a noose shaped like the number six.
 - Relevance:** This document is directly related to a specific event and the broader campaign against apartheid-related executions. It can provide a case study on how activists mobilized international support to oppose the apartheid government's use of the death penalty against political activists, demonstrating the power and strategies of the anti-apartheid movement.
- Isolate apartheid: sanctions now, victory to the ANC:**
 - Content:** Contains photographs of children and adults rallying with the ANC flag and posters, including images of Nelson Mandela.
 - Relevance:** This poster offers insights into the role of the African National Congress (ANC) and the call for economic sanctions against South Africa. It highlights the internal struggle within South Africa and the international campaign to isolate the apartheid regime economically and politically. The mention of Nelson Mandela also opens avenues for discussing key figures in the fight against apartheid.

These documents can collectively help illuminate the multifaceted nature of apartheid, the various forms of resistance against it, and the global efforts that contributed to its eventual dismantlement. They provide a mixture of visual and textual resources that can be analyzed for their historical, cultural, and political significance in the context of apartheid and its legacy.

Figure 2. This is another example interaction from NUL's current chat application. In this example, the response from the LLM describes the content of each returned result and its relevance to the user's question.

The OSDP will be based on the work done for NUL's current chat application. It will be 'one click' installable in Amazon Web Services using an abstracted cloud deployment method. The initial product will include all necessary components of a RAG-based solution including vector storage, processing and embedding, and routing, as well as configurable integration with AI endpoints, such as OpenAI's ChatGPT4. The OSDP intends to provide a way for library staff to become more informed about RAG solutions that are going to be offered by library vendors like ExLibris and OCLC. Libraries "need to be equipped as savvy buyers of AI tools. They need to understand how biases manifest in these tools so they can experiment while maintaining their organization's ethical standards."⁶ NUL will host a workshop to teach other libraries about this product, its architecture, the strengths and weaknesses, and how it might be deployed in other contexts.

⁶ <https://hai.stanford.edu/sites/default/files/2024-02/Working-Paper-Inspiring-Action.pdf>

The Need, Deliverable Two: Metadata Toolkit

Leveraging the RAG-based architecture in the OSDP, the project's separate metadata toolkit will showcase an end-to-end AI assisted metadata workflow for digitized collections. Large-scale digitization projects often operate with limited staff meaning metadata remains high-level and not specific to each item. Details are missed or overlooked causing works to be published without necessary context.⁷ Additionally, most of these projects are born out of archival collections that subscribe to the "more product, less process" approach so that the descriptions generated as part of processing are also often broad and definitely do not describe at the item or even folder level.⁸ Item level descriptions of digitized works will increase more precise access to individual works.

The toolkit will allow technical users to experiment with generative tools to augment human workflows in assigning Library of Congress (LOC) subject headings at the item level and writing item level image descriptions using a multimodal approach. As seen in Figure 3, NUL created an experimental application that shows how AI can be used to augment human workflows for creating item level description. Tools for experimentation and analysis of images for racist, biased, or otherwise potentially harmful content will be built into the toolkit.⁹ This experimentation work will view generative AI as an opportunity "to identify and reduce the effect of human biases," while also documenting any biases the chosen LLM seems to perpetuate and starting the work necessary to fine-tune and improve these systems as they are deployed.¹⁰

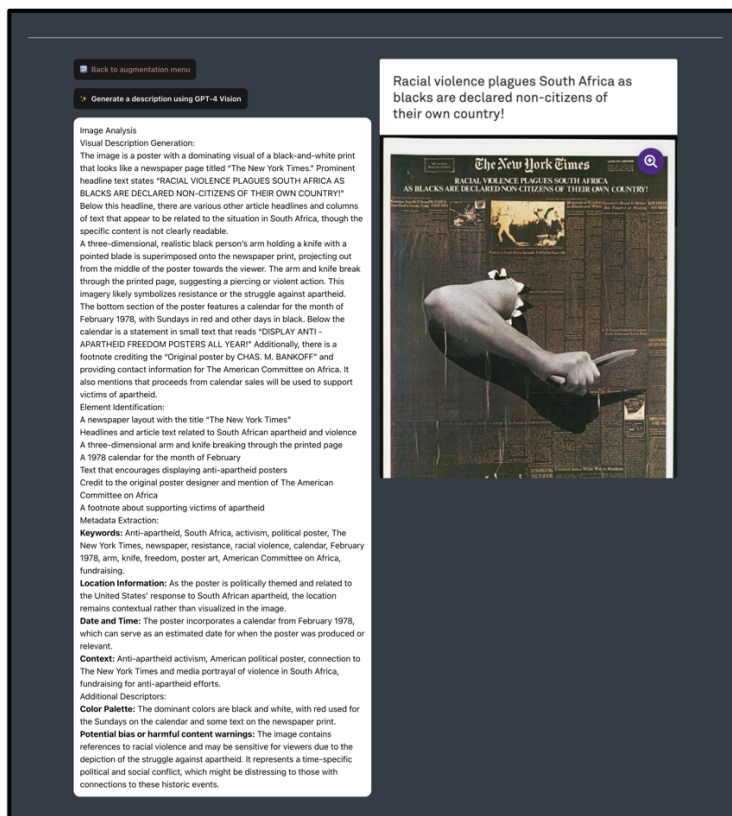


Figure 3. NUL experimental application showing descriptive capabilities of multimodal GPT-4.

⁷ Lapworth, Emily; Jones, Sarah; and Georgieva, Marina (2019) "Microfilm, Manuscripts, and Photographs: A Case Study Comparing Three Large-Scale Digitization Projects," *Journal of Contemporary Archival Studies*: Vol. 6 , Article 5.

⁸ <https://elischolar.library.yale.edu/jcas/vol6/iss1/5>

⁹ <https://meridian.allenpress.com/american-archivist/article/68/2/208/24011/More-Product-Less-Process-Revamping-Traditional>

¹⁰ <https://aqlanthology.org/2022.findings-acl.165.pdf>

¹⁰ <https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>

The end-to-end AI assisted metadata workflow will be implemented, tested, and iterated on in NUL's Digital Collections backend. By solving real-world metadata problems in a production environment, this will ensure adopters learn not only with a "toy" solution, but can verify the usefulness in a public, large scale, highly trafficked digital collection site. The workshop NUL will host for other libraries will also include a deep dive into the metadata toolkit and how other libraries may apply these methods within their own institutional contexts.

Target Audience and Beneficiaries

The target group for the OSDP and metadata toolkit are libraries that have an interest in learning how these new generative AI technologies work in the discovery setting and how they can be used in creating metadata. NUL has given multiple presentations about its Digital Collections chat-based search approach, including at Samvera Connect 2023 and the BTAA Library Digital Strategies and IT Services group, which generated follow up and interest. Consequently, NUL created a workshop for 2024 Code4Lib, which has sold out. NUL was asked to informally consult with libraries, such as Boston Public Library and Princeton University Library, about our work in this area. Because of this interest and demonstrated need, we plan to host a workshop as a key part of this project for 10-15 libraries to get started with implementation of this discovery chat-based product and metadata toolkit, gather feedback for improvement, as well as create well-designed online resources for other libraries to get started asynchronously. In terms of beneficiaries of this project, it will be the end users of library discovery systems that will benefit from this new way of discovering information and that will benefit from libraries using generative AI as a co-pilot to create more metadata for their digital collections. As a starting point, NUL's Digital Collections averages more than 150,000 sessions a year, which will provide a well trafficked website with a global reach, including Africa, Asia, South America, and Western Europe, for testing the project's deliverables.

Building Upon and Complementing Current Landscape

This proposed project complements other existing products that leverage Retrieval Augmented Generation (RAG) architecture and best-in-class generative AI models, such as the open source WARC-GPT product from Harvard's Law Library Innovation lab.¹¹ Harvard created that product for web archives and our OSDP is designed for digitized collections and metadata sets from traditional library resource catalogs. In terms of using generative AI to assist in creating and augmenting metadata, we have instead found examples of projects using traditional machine learning. For example, OCLC has used machine learning techniques to remove duplicate records, but is different as it proposes to use generative AI tools.¹² An example that complements this project is the Audiovisual Metadata Platform, but AMP focuses primarily on describing audiovisual materials, which pose unique access and preservation issues.¹³ Instead, our project will largely focus on describing image-based collections and providing a blueprint for generative discovery. The Audiovisual Metadata Platform provides a template for how to design documentation so that other libraries can adapt the metadata toolkit we propose.

The Advisory Board, made up of a diverse set of library administrators and managers with expertise around AI, is another key aspect of this project. The Board will provide a sounding board to ensure the project complements existing products as it progresses (see Diversity Plan for names of confirmed members). NUL designed this project to not only create tangible deliverables, but to also empower libraries to learn about and

¹¹ <https://lil.law.harvard.edu/blog/2024/02/12/warc-gpt-an-open-source-tool-for-exploring-web-archives-with-ai/>

¹² <https://hangingtogether.org/machine-learning-and-worldcat-improving-records-for-cataloging-and-discovery/>

¹³ <https://uisapp2.iu.edu/confluence-prd/display/AMP/AMP%3A+Audiovisual+Metadata+Platform>

effectively utilize generative AI technology to address the challenges we face in providing access to information and maintaining collections in this new world.

2. Project Work Plan

The proposed work plan encompasses the two main deliverables, documentation, and an associated workshop. **Deliverable One:** the release of the installable OSDP. **Deliverable Two:** The publication of the metadata toolkit for digitized collections. For both deliverables, NUL plans to leverage and iterate on existing work, which will improve access to its Digital Collections while generating interest in methods for adoption at multiple libraries. With the workshop and asynchronous training materials, we look toward a commitment of adoption in at least 10-15 libraries by grant end.

Key People and Methods: Iterative goals and work will be planned and organized by David Schober, Team Lead and Products Manager. David Schober has led successful large-scale grant and internally funded projects at Northwestern (Avalon, Digital Collections, Meadow, Serverless-IIIF) as well as projects with large, global reaches at other institutions including the American Academy of Pediatrics (Pediatrics journals-suite redesign), and the Poetry Foundation (Digital Transformation). Senior Developer Brendan Quinn will take a leadership role in technical implementation and developer toolkits and work extensively with a Senior Developer to be hired with grant funds. Brendan has worked at Northwestern for 14 years, recently serving on the Provost-led Generative AI Advisory Committee.¹⁴ He has spoken widely about the intersection of generative AI and libraries.

NUL manages its projects in the open and uses best practices from Scrum and other agile methodologies.¹⁵ All work is visible to the team, management, and stakeholders via GitHub issues and open presentations (demos). Work is broken out into high-level goals and the team working on the project focuses on iterative deliverables in eight-week cycles. In addition to the project director and project manager, the project team will consist of cross-functional developers, metadata subject matter experts (Jamie Carlstone and Jen B. Young), user experience professional (Frank Sweis), and communication professional (Cory Slowik). Throughout the cycle work on deliverables one and two will occur concurrently as the team works in small iterative cycles.

Deliverable One Goals: Semantic Search Iteration and Generalization (Months 1-8, 14-24). The team will iterate on our initial public release of the Digital Collections generative semantic search. In collaboration with our User Experience librarian, we will collect user feedback using various descriptive, evaluative, and analytic research methods. These may include contextual inquiries and interviews, focus groups, usability studies, quantitative data analysis from actual use (analytics and anonymized queries), and longitudinal activities such as diary studies. The team will also apply other machine learning techniques to augment the results and improve results.

From the Digital Collections platform, we will generalize the technology to create the OSDP into a deployable Web application and associated APIs. Libraries will be able to install the OSDP, load content, generate embeddings, and perform chat-based search using industry-leading LLMs. The OSDP will allow libraries to experiment and evaluate various LLMs and their output on their own collections. For instance, libraries will be able to configure embedding models to evaluate either commercial solutions or open-source foundational models. GitHub repositories containing all production source code, documentation, and example python notebooks will be released and updated throughout the grant. In addition, a “brochure” style website will be produced that acts as a hub for the project.

¹⁴ <https://www.northwestern.edu/provost/about/committees/generative-ai-advisory-committee/>

¹⁵ <https://github.com/orgs/nulib/repositories>

Deliverable One Key Activities:

1. Experiment with several commercial and opensource LLMs (Llama, Claude, ChatGPT) and build abstraction tooling so implementers can evaluate embeddings and generation. (M1-M3 Schober, Senior Developer, Quinn)
2. Experiment with multimodal embedding of image, video, and audio objects. (M1-M3 Schober, Senior Developer, Quinn)
3. Develop UX tests that validate LLM responses for accuracy and validity. (M4-M9 Schober, Sweis)
4. Release Digital Collections semantic generative search for public-use (move from campus-access-only beta). (M7-M9 Schober, Quinn, Sweis)
5. Generalize and extract the embedding and search pipeline for use in the OSDP. (M5-M8 Schober, Senior Developer, Quinn)

Deliverable Two Goals: Metadata Generation and Validation (Months 9 –14, 14-24). The team will develop open-source software for use in Northwestern’s preservation DAM (Meadow) that augments human metadata workflows with LLMs (initial target ChatGPT) to describe objects and add subject headings from standard controlled vocabularies, including Library of Congress (LoC) and Getty. The development team will explore various methods of prompting and fine tuning to improve automated responses and results. The development team, with the user experience librarian and metadata librarians, will define and design solutions for common bottlenecks in workflows when describing digitized collections at scale. The team will identify and document pitfalls and significant improvements to workflows. Finally, the team will publish the toolkit for other libraries as a methods document and a playbook with examples and code samples for implementing similar pipelines.

Deliverable Two Key Activities

6. Investigate best models to use for description and subject headings (M10-M12 Schober, Senior Developer, Quinn).
7. Experiment with prompting methods (M10-M12 Schober, Senior Developer, Quinn).
8. Embed and populate vector databases with descriptions subject headings including LoC and Getty (M13-M15 Schober, Senior Developer, Quinn, Young, Carlstone).
9. Develop methods of validation and reinforcement against domain-specific information and knowledge (M13-15 Schober, Senior Developer, Quinn, Young, Carlstone).
10. Test workflows with subject matter experts, refine methods based on feedback (M16-18 Schober, Senior Developer, Young, Carlstone).
11. Test AI descriptions at scale with NUL Berkeley Folk Music collection, a large multi-faceted collection with a well-documented workflow and hours to complete (M16-18 Schober, Senior Developer, Quinn, Young, Carlstone, Caizzi).
12. Publish Metadata Toolkit document methods, strengths, and shortcomings of using AI to augment description at scale (M19-M20 Sweis, Senior Developer, Quinn, Carlstone, Young).

Related Workshops, Education, and Adoption:

Throughout the work on the OSDP and the Metadata Toolkit, the team will develop education and adoption materials, notably through the previously mentioned workshop. In addition to methods and theory, the workshop will focus on installing and using the OSDP with data from the attending institution’s digital collections. During the workshop the metadata toolkit will be introduced, and users will learn through hands-on use of various aspects of LLM-assisted metadata creation examples. At the same time, we will develop workshops

for submission to community conferences. The team's success with various open-source projects highlights the critical aspect promotion and training has on the success of open products. The in-person workshop will serve to educate adopters and grow the library community's knowledge in this critical field and act as a feedback loop for iterations on the released open-source tools.

13. Develop workshop content for installing, creating vectors, querying and customization using the OSDP and metadata toolkit. (M16-M18, M22-M24 Schober, Sweis, Senior Developer, Quinn, Slowik)
14. Host 2-day workshop at NUL for 10-15 library staff including IT professionals and metadata librarians (M23 Schober, Caizzi, Quinn, Senior Developer, Carlstone, Sweis, Young).
15. Develop workshops for community-led conferences (Samvera Connect, Code4Lib, etc.) (M20-M22 Quinn, Senior Developer, Carlstone, Young, Caizzi, Schober, Slowik).
16. Develop website and related marketing materials for the OSDP (M13-M15 Slowik, Schober, Sweis).

3. Diversity Plan

The diversity plan for this project focuses on 2 distinct areas: 1) end user accessibility for OSDP and 2) incorporating a diversity of perspectives throughout the iteration of both deliverables (OSDP and metadata toolkit).

In terms of end user accessibility, on-the-fly translation is a feature of contemporary large language models. Initially trained on English-language sources, the updated best-in-class solutions now offer powerful translation and comprehension in many languages. While acknowledging there are issues with translations, being able to offer natural language search in languages beyond English moves us in the right direction—offering users far more than a library could reasonably staff and provide. For example, see the screenshot from our OSDP prototype of a complex search in Italian, created and verified by a native Italian speaker (Figure 4). Additionally, the metadata toolkit will experiment with generating item level descriptions of digitized collections. A lack of thorough item level descriptions in digitized collections may act as an obstacle for visually impaired users who often must settle for overly general descriptions of visual works. The metadata toolkit would showcase how these tools can be used cost-effectively to construct ALT tags fully describing images (see Figure 3 in Project Justification as an example of what information can be included in an ALT tag).

Throughout the project we plan to utilize user studies and a diverse set of testers to validate the above and drive iterations. Our goal is to validate and test with users LLM-assisted search and metadata creation in a variety of contexts. These contexts will be managed by the user experience librarian, who will work to identify NUL staff, Northwestern students and faculty to take part in providing user feedback. For example, NUL has a student advisory board that is intentionally diverse from undergraduates to graduate students from a wide-ranging set of majors and the User Experience librarian plans to ask them to participate in our user studies. Additionally, the Advisory Board members are made of diverse set of academic libraries which will serve as a guidepost for ensuring that the needs of institutions like theirs are met. The user experience librarian will conduct the two feedback sessions with the Advisory Board and that feedback will be incorporated into our sprints for both deliverables. The confirmed participants for the project Advisory Board are:

- Bohyun Kim, Associate University Librarian, University of Michigan
- Cathryn Steeves, Vice-Provost and Chief Librarian, Western University
- Lauren Di Monte, Associate University Librarian for Teaching and Research, Yale University
- Leo Lo, Dean and Professor, University of New Mexico
- Esmé Cowles, Assistant Director of Library Software Engineering, Princeton University
- Tim McGeary, Associate University Librarian for Digital Strategies and Technology, Duke University

- Elizabeth Gushee, Associate Dean Digital Strategies and Scholarly Communication, University of Miami
- Tom Cramer, Associate University Librarian Digital Library Systems and Services, Stanford University
- Catherine Brobson, Product Manager, Boston Public Library
- Stephen Downie, Associate Dean for Research, and Co-Director HathiTrust Research Center, University of IL at Urbana-Champaign

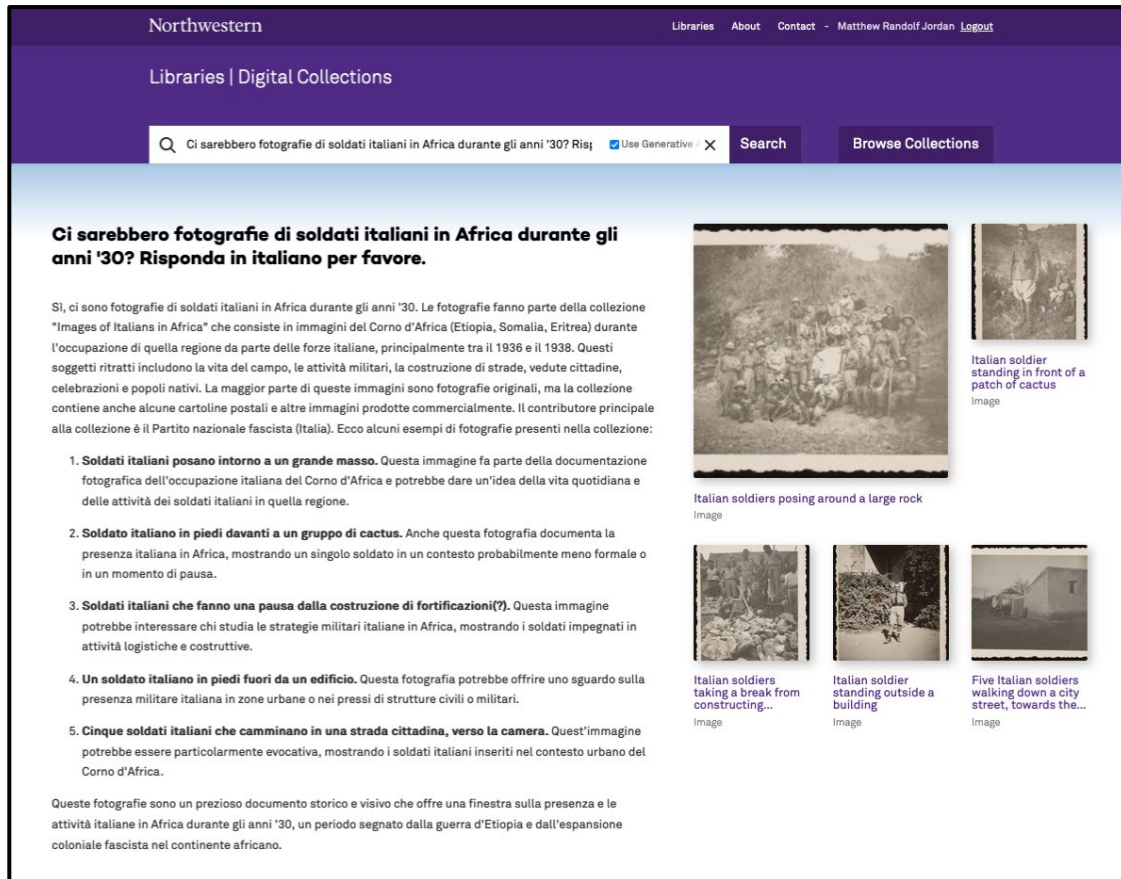


Figure 4. NUL's current chat application allows users to ask questions and receive replies in foreign languages. In this example, the LLM returned valid responses along with context in the language the user requested a response in.

4. Project Results

The project's intended results are to create two deliverables, the OSDP and metadata toolkit as described in the Project Justification. With these results, libraries will be able use the power of generative AI for innovating new ways to approach collection management meeting the new expectations users have in terms of finding information online.¹⁶ Additional results include:

- leveling up library staff knowledge and expertise from a variety of institutions
- growing skills to assess new AI products from library vendors and other commercial vendors
- adoption of the OSDP and metadata workflows learned from the toolkit in 10-15 libraries

¹⁶ See footnote 5. Here is another market analysis of changing user expectations: <https://www.tidio.com/blog/year-of-chatgpt/>

As Lorcan Dempsey states “Libraries will need to become knowledgeable about claims about AI, functionality, uniqueness - as they are in other areas of procurement.”¹⁷ In sum, this project will advance libraries understanding of this new technology, allow them to have some control over their discovery tools and management of metadata, and will benefit end users by advancing knowledge.

NUL will ensure a usable open-source discovery product (OSDP) in multiple ways. First, NUL will release a testable version in AWS Web Services, which is used by tens of thousands of educational institutions worldwide.¹⁸ Making this platform installable using a service such as Amazon’s Serverless Application Repository, adopters have immediate access to a globally redundant, production-ready environment at a minimal cost. Secondly, since NUL is very active in library open-source and open standards communities, such as IIF and Samvera, we have developed key relationships with other academic libraries who adopt and contribute back to open-source products, which will provide scaffolding for the sustainability and usability of a product like this. For example, NUL partnered with Indiana University to develop and release the successful open-source product, Avalon Media System, that is in production in dozens of libraries. Another example that ensures NUL has experience creating adaptable and sustainable open-source products are Serverless-IIF, Canopy, and the Clover IIF viewer, all of which were developed initially by Northwestern’s team, generalized, then adopted by a wider community.¹⁹ In terms of the metadata toolkit, the metadata specialists will also rely on the community of digital collections professionals from various communities, such as the Digital Library Forum community, to ensure that the toolkit is usable beyond NUL. Lastly, the Advisory Board of AI library experts mentioned in the Diversity Plan for the project will function as a sounding board and feedback loop to ensure both the discovery product and metadata toolkit are usable in the variety of libraries they represent.

As stated above, NUL is no stranger to sustainability of open-source projects and products. Because NUL will be running the end user chat product in production at the start of this grant project and will encourage community contributions to OSDP during the grant project, its development will not cease after grant funding. NUL hopes to add the grant funded developer to recurring funds at the end of grant period since the project will demonstrate a business need. The current team will continue to prioritize this product as NUL sunsets older products if the grant funded developer cannot be retained after grant period. Another aspect of sustainability is the approach to marketing and communication about the product. In addition to its availability for testing in AWS, like Serverless IIF, it will have an accompanying GitHub repository as well as corresponding technical and user documentation. An example of NUL’s approach to open-source product management in terms of documentation includes Canopy mentioned earlier, a static site generator for IIF resources. NUL's strategic approach to sustainability for products is to make multiple conference presentations, make connections with other developers, and create a robust technical and end user documentation site, which drives contributions back into its codebase on GitHub. In sum, NUL has a strong history of sustainability of its open-source projects and products beyond grant funding and will treat both grant deliverables no differently.

¹⁷ <https://www.lorcandempsey.net/generative-ai-and-libraries-7-contexts/>

¹⁸ <https://aws.amazon.com/education/>

¹⁹ <https://samvera.github.io/serverless-iiif/>; <https://canopy-iiif.github.io/docs/>; <https://samvera-labs.github.io/clover-iiif/>

Year 1

M=Month; A=Activities	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
A1: LLM experimentation.												
A2: Multimodal embedding.												
A3: UX Tests and Validation.												
A4: General Public Release of chat-based search.												
A5: Extraction and generalization.												
A6: Investigate best models to use for description and subject headings.												
A7: Experiment with prompting methods.												

Public Release of Digital Collections Semantic Search

Northwestern will open the semantic search tool for general public consumption on [Digital Collections](#). The release will incorporate several iterations of feedback during our pilot and beta launches. The semantic search will offer users a new way to interact with Digital Collections that offers descriptive context to complex queries, and creative generative answers to open-ended questions. In addition, the public release will focus on smooth integration with existing search expectations offering users unfamiliar with semantic and generative search tools an “on-ramp.” Images, video, and audio on Digital Collections include materials in the public domain and those under copyright and made available under fair use or donor agreements. All source code, documentation, and related material will be released under the [Apache-2.0 License](#). Users are permitted to modify, distribute, and sublicense the original open-source code and content. Northwestern approaches its software development through a sustainable, continuously iterated product. The product will be sustained through NUL’s commitment to continuous development. Many examples of its commitment to other locally developed products are in the narrative.

Chat-based Discovery Product (OSDP)

The project will generalize work in progress at Northwestern in LLM-based semantic search for delivery as a stand-alone chat-based semantic search tool. The tool will act as a starting point for other institutions looking to integrate chat-based search into user-facing digital collections and library catalogs. The product will be deployable for use in a cloud environment via the Serverless Application Repository (AWS) and will rely on public APIs for LLM access (Azure, OpenAI) as well as an OpenSearch vector store and related front-end based on ReactJS. This will allow for a “one-click” installation of the environment for experienced and novice experimenters. Once deployed, implementers will be able to create vectorized data out of existing metadata and query the results with natural language. The product will not record private information, use of third-party APIs is subject to terms of use. Azure’s OpenAI endpoint ensures [security and privacy](#) in terms of service. In addition, all source code, documentation, and related materials will be available for review, customization, and pull-requests. All source code, documentation, and related material will be released under the [Apache-2.0 License](#). Users are permitted to modify, distribute, and sublicense the original open-source code and content. Northwestern approaches its software development through a sustainable, continuously iterated product. The product will be sustained through NUL’s commitment to continuous development. Additionally, NUL will encourage participation and community contributions. Many examples of its commitment to other locally developed products are in the narrative.

Metadata Generation and Validation Toolkit

The project will develop a metadata generation and validation toolkit that augments human metadata workflows with LLMs to describe objects and validate against standard controlled vocabularies, including Library of Congress (LoC) and Getty. The toolkit will be comprised of example code, training materials, Jupyter notebooks, and purpose-built software libraries for abstraction of common patterns. Web-based tools will be deployed to demonstrate metadata generation, validation, and remediation. All source code, training materials, and documentation will be available publicly on Northwestern’s GitHub repository. All source code, documentation, and related material will be released under the [Apache-2.0 License](#). Users are permitted to modify, distribute, and sublicense the original open-source code and content.

Promotional Materials

A suite of promotional materials will be created as part of this project including but not limited to: a brochure-style site for marketing the chat-based discovery product to new users; documentation and technical resources about installing, using and implementing the system on production environments; a brochure-style site for the metadata generation toolkit and suite of educational materials; blog posts and other ancillary materials. All content and related source code will be released under the [Apache-2.0 License](#). Users are permitted to modify, distribute, and sublicense the original open-source code and content.