NATIONAL LEADERSHIP GRANTS FOR LIBRARIES

Sample Application LG-259840-OLS-26
Project Category: Applied Research

Virginia Tech University Libraries

Amount awarded by IMLS:	$473,403
Amount of cost share:	$0

Virginia Tech University Libraries will investigate how artificial intelligence-generated metadata in digital libraries is represented. The project seeks to determine whether large-scale machine-generated metadata can serve as reliable infrastructure or its variability introduces limitations that must be understood and overcome before it is incorporated into patron-facing digital library systems. The intended project results are to establish empirical evidence about when machine-generated metadata can function as a structural component of digital library systems and when it cannot. Specifically, the project will identify which categories of AI-generated metadata remain sufficiently consistent when produced across large corpora, analyze the structural relationships that emerge across collections, and evaluate how these relationships influence discovery and interaction with digital materials. The goal is to provide libraries with a research-based foundation for determining whether and how large-scale AI-generated metadata can be incorporated into digital library infrastructure without undermining the principles of reliability, transparency, and long-term stewardship that govern professional practice.

Attached are the following components excerpted from the original application.

- Narrative
- Schedule of completion

When preparing an application for the next deadline, be sure to follow the instructions in the most recent Notice of Funding Opportunity for the grant program to which you are applying.

Library-wide, Machine-Generated Metadata as First-Class Structural Components

Bipasha Banerjee, Jennifer L. Goyne & William A. Ingram

INTRODUCTION

Virginia Tech University Libraries seeks IMLS support of $519,962 for a three-year applied research grant titled *Library-wide, Machine-Generated Metadata as First-Class Structural Components* to gain insight into the characteristics, limitations, and implications of AI-assisted metadata generation across heterogeneous digital collections. Digital libraries depend upon descriptive metadata to render collections intelligible and navigable across time and institutional contexts. For most of their history, this descriptive layer has been produced through human cataloging, constrained by labor, expertise, and institutional priorities. As a result, digital library systems have been designed around an implicit condition of metadata scarcity. Recent advances in AI challenge this assumption. It is now technically possible to generate descriptive information across entire textual corpora, producing entities, relationships, and contextual signals at scales that were previously unattainable through human cataloging alone. However, the large-scale generation of machine-produced metadata also raises concerns about reliability, consistency, and authority. Metadata generated automatically across large collections may introduce redundancy, instability between model runs, or conflicts with established descriptive practice. The emergence of such corpus-scale machine-generated metadata raises a foundational question for the field, whether any forms of this metadata be produced with sufficient consistency, reliability, and structural coherence to function as first-class components of a digital library. Our research focuses on three questions central to digital libraries:

- **RQ1:** Which AI-generated metadata can be produced across collections with sufficient consistency to function as structural components of a digital library system?
- **RQ2:** What structural relationships and entities emerge in digital libraries when machine generated metadata is abundant at corpus scale?
- **RQ3:** What new forms of person-centric, event-centric, community-centric, and cross-format interaction become feasible when structural relationships and entities emerge in machine generated metadata?

By examining these issues within an operational digital library environment, the project seeks to determine whether (a) large-scale machine-generated metadata can serve as reliable infrastructure or (b) its variability introduces limitations that must be understood and overcome before it is incorporated into patron-facing digital library systems. Our intended project results are to establish empirical evidence about when machine-generated metadata can function as a structural component of digital library systems and when it cannot. Specifically, the project will identify which categories of AI-generated metadata remain sufficiently consistent when produced across large corpora, analyze the structural relationships that emerge across collections, and evaluate how these relationships influence discovery and interaction with digital materials. These findings will provide libraries with a research-based foundation for determining whether and how large-scale AI-generated metadata can be incorporated into digital library infrastructure without undermining the principles of reliability, transparency, and long-term stewardship that govern professional practice. In addition to advancing scholarly understanding of AI within digital library architectures, the project will produce openly documented methods, software prototypes, and guidance materials that allow other institutions to examine similar questions within their own collections and systems. The research and its focus on the use of AI in library services align with the federal AI strategy for 2025 (Executive Office of the President, 2025), by advancing the scalable integration of AI within libraries.

PROJECT JUSTIFICATION

Needs, Challenges, and Opportunities

Digital libraries function as core national infrastructure for research, scholarship, cultural memory, and public knowledge in the United States. For years, librarians and archivists have relied on systems of

description to render collections intelligible, navigable, and accountable over time. Digital libraries in use today were largely designed in an era when metadata was manually authored and therefore limited by the availability of professional labor and domain expertise. Current systems for discovery and navigation reflect this historical context.

The conditions under which libraries produce metadata are rapidly changing. Generative AI has made large-scale metadata creation less resource-intensive, although ongoing research must continue to address questions of consistency and reliability. Metadata has long functioned as a structural layer within digital library systems, supporting description, organization, discovery, and retrieval of materials. Digital libraries now confront a different condition in the age of AI, the possibility of metadata abundance. Cross-collection linkages and normalization, automated relationship mapping, and fine-grained information extraction are some of the possibilities that arise when libraries can use AI to generate metadata at scale. The situation creates an opportunity to experiment with how large-scale AI-generated metadata might reshape structural representations, reveal relationships, and alter how discovery and retrieval operate at the level of the collection as a whole.

Prior research in digital libraries and information retrieval has explored techniques for extracting descriptive information from textual collections (Banerjee et al., 2024a; Jadhav et al., 2025). Investigations into entity recognition, semantic indexing, knowledge graph construction, and more recently large language model (LLM)–assisted metadata generation have largely examined the performance of such techniques in discrete contexts, such as improving retrieval rankings, identifying entities, or generating descriptive annotations for individual items. However, digital libraries do not operate as collections of isolated extraction tasks; they function through the cumulative interaction of descriptive structures that organize collections, normalize relationships among digital objects, and sustain meaningful connections through cultural and scholarly records. When machine-generated metadata is produced across entire collections rather than for individual items, it ceases to be merely descriptive output and begins to operate as part of the structural fabric of the system itself. The consequences of this shift remain largely unexplored.

This proposal investigates how machine-generated metadata can function as architectural components of production digital libraries. The research will be conducted using collections hosted within the Virginia Tech University Libraries Digital Library Platform (DLP) (VTUL, 2024), a production environment that hosts a wide variety of materials. DLP collections include oral histories, community archives, regional organizational records, and digitized cultural heritage materials documenting rural Appalachian communities, and reflect the uneven description of many rural archival holdings produced under conditions of limited resources. This context for evaluating an appropriate empirical setting for research into how large-scale AI-generated metadata might alter the structural representation and enable cross-collection linkages across collections that are organizationally dispersed but geographically and historically connected. By producing openly documented methods, reference architectures, and technical guidance for digital library reseachers and practitioners, the project contributes to the ability of American research libraries to incorporate computationally derived metadata structures into systems responsible for stewarding complex digital collections.

Relationship to Existing Models, Standards, Scholarship, and Practice

Digital library systems rely on structured descriptive metadata fields such as title, subject, and description to support the organization and retrieval of materials. Lu et al. (2008) provide an early benchmark of structural multi-leveled metadata from OCR'd journals. The authors emphasize that metadata should be viewed as the foundational infrastructure and demonstrate how it supports navigation within an operational digital library. Fielded retrieval models such as BM25 (Robertson et al., 2004) demonstrate how metadata fields influence ranking behavior within digital library systems by weighting signals from different parts of the record. Because these descriptive fields play a central role in discovery, digital library research has increasingly explored computational methods for extracting metadata, entities, and relationships from cultural heritage collections. Ruotsalo and Hyvönen (2007) propose an event-based knowledge schema that

uses domain ontologies to integrate heterogeneous metadata formats across systems. This work led to the development of the "CultureSampo" semantic portal. Such systems demonstrate how relational metadata can reorganize cultural heritage materials and support new modes of discovery. However, these approaches typically rely on curated ontologies and manually constructed entity relationships rather than metadata generated automatically across entire collections.

Recent work has shown that models can generate entities, relationships, and descriptive signals from collections. Wu et al. (2020) present a scalable BERT-based (Devlin et al., 2019) entity-linking retrieval approach that outperformed BM25 on three datasets. El Ganadi et al. (2025) integrate OCR and LLM-assisted metadata extraction of their Arabic digital library. The work indicates the feasibility of the approach but also highlights the challenges posed by the language. Arnold and Tilton (2024) suggest using multi-modal LLMs for images by creating textual descriptions and text-based embedding. Bagchi (2024) describe a human-AI collaboration approach for restructuring library metadata across five representation layers. Keraghel et al. (2024) conduct a comprehensive study of named entity recognition (NER) approaches that use LLMs, graph-based approaches, and reinforcement learning techniques. Their findings suggest transformer-based models such as GliNER (Zaratiana et al., 2024) perform well on NER tasks across standard NER datasets. The GPT-NER (Wang et al., 2025) framework converts NER into a generation task and using in-context learning for the LLM. These AI methods show that metadata, entities, and descriptive text can be automatically generated from library collections, including OCR-derived and multimodal materials. What remains underexplored is whether the resulting metadata is consistent and stable enough to function as structural infrastructure, rather than as task-specific output.

AI-enabled metadata management has also been explored. Oyighan et al. (2024) discusses some key challenges and hurdles in maintaining high quality and consistency while emphasizing the importance of integrating AI into digital libraries to significantly help discoverability. Yang et al. (2025) examine how metadata management is shifting from manual to AI-enabled processes across open-source, proprietary, and research systems. Protocols on using AI and integrating into digital libraries and archives are being developed (Lo, 2026). Libraries are already moving into AI-enabled metadata environments, and the profession is beginning to articulate concerns about quality, consistency, and governance.

Existing work demonstrates that machine learning and language models can extract entities, relationships, and descriptive signals from digital collections, and that structured metadata and knowledge graphs can organize and link cultural heritage materials across systems. However, these studies typically evaluate extraction methods or retrieval applications rather than examining how machine-generated metadata behaves once it becomes widely integrated enough serve as structural infrastructure within digital library systems. The architectural implications of large-scale machine-generated metadata, such as its consistency, stability, and influence on discovery, remain largely unexplored. This project addresses that gap.

The Benefit to Libraries

This applied research provides an empirical basis for evaluating whether AI-generated metadata can function as reliable structural components within digital library systems and how such structures should be integrated into library infrastructure. The project advances the Institute of Museum and Library Services National Leadership Grants for Libraries (NLG-L) Objective 1.3 by strengthening the information infrastructures through which libraries preserve collections and provide broad access to collections, while supporting new forms of exploration across digital collections.

Libraries must understand how AI-generated metadata should function within the descriptive and structural infrastructure of digital collections. Cataloging standards, authority control, subject vocabularies, and metadata schemas are mechanisms through which the profession governs knowledge organization. AI-generated metadata can extend this infrastructure by identifying entities and relationships across entire corpora. If such metadata can be produced consistently at corpus scale, libraries could represent people, communities, events, and relationships across collections in ways that extend beyond traditional document-

level description, enabling discovery organized around entities and connections that span multiple works and formats.

At the same time, the rapid adoption of AI-driven research platforms raises the risk that structural metadata will be generated and controlled by systems whose methods and assumptions are opaque to libraries. If large-scale structural metadata is produced by such platforms and libraries lack the ability to evaluate, interpret, or design those systems, the profession risks losing its role as the steward of descriptive infrastructure. Without research that examines how machine-generated metadata behaves at corpus scale and how it can be incorporated into digital library systems, libraries may become dependent on proprietary infrastructures that shape how collections are interpreted and discovered.

PROJECT WORK PLAN

Our project is divided into three phases with each phases roughly spanning a year. Table 1 shows each of the phases, research questions and associated tasks.

Table 1: Research phases, questions and associated tasks

Phase	Research Question	Tasks
Phase 1	**RQ1:** Which AI-generated metadata can be produced at corpus scale with sufficient consistency to function as structural components of a digital library system?	1. Data Selection 2. Experimental Workflow 3. Scaling and Evaluation
Phase 2	**RQ2:** What structural relationships and entities emerge in digital libraries when machine generated metadata is abundant at corpus scale?	4. Build Structural Representations 5. Analyze Cross-collection Entities 6. Event-Temporal Linkage Analysis
Phase 3	**RQ3:** What new forms of person-centric, event-centric, community-centric, and cross-format interaction become feasible when structural relationships and entities emerge in machine generated metadata?	7. Develop Navigation Prototype 8. Cross-collection Discovery Analysis 9. Feasibility Assessment of Generated Narratives

Phase 1

Phase 1 investigates Research Question 1 by evaluating whether AI-generated metadata can be produced with sufficient consistency to function as structural components of a digital library system. It begins with the construction of an experimental dataset drawn from the diverse collections hosted in the Virginia Tech Digital Library Platform (DLP) (VTUL, 2024). A representative subset of collections is selected to provide a range of formats, content types, and descriptive conditions suitable for evaluating AI-generated metadata at scale. Using this dataset, we establish a controlled experimental workflow for generating and evaluating AI-derived metadata. The workflow builds upon our prior work on text extraction and metadata generation within DLP collections (Banerjee et al., 2024a; Jadhav et al., 2025).

We evaluate several open-weight LLMs under controlled prompting and parameter settings to measure the consistency of generated metadata. Once we have a set of parameters established for consistency, we move on to the next task for scaling and evaluation. In this study, consistency refers to the stability of generated metadata across repeated model runs, the agreement between AI-generated metadata and existing human-authored fields when available, and the persistence of extracted entities in documents within the collection. For scaling, we want to find the boundaries of which metadata fields can be scaled

to a corpus level with consistency. We conduct behavioral testing on the selected models to check for prediction correctness and robustness.

Task 1: Data selection. We construct a dataset drawn from collections hosted in the DLP, which currently contains twenty-three collections totaling more than 57.6 TB of digital objects. From these holdings we select a representative subset of collections that reflect the diversity of materials described earlier in this proposal, including oral histories, community archives, regional organizational records, and digitized cultural heritage materials. These collections span multiple formats and modalities such as PDFs, JPEG images, TIFF scans, and audiovisual materials that contain textual and multimedia content. The subset is chosen to include collections containing a broad range of entities such as people, places, events, and communities so that entity extraction and relationship identification can be evaluated across heterogeneous materials. Machine-readable text derived from the collections, along with existing metadata, will be used to cluster entities present in the dataset. The selection process and resulting dataset characteristics will be documented to support reproducibility and analysis of the experimental results.

Task 2: Experimental Workflow. To develop the experimental setup, we start by identifying metadata fields that can be scaled to a collection and corpus level with sufficient consistency. Candidate fields include common descriptive, structural, and administrative metadata elements used in digital library systems.

Once candidate metadata fields have been identified, the experimental workflow proceeds in three steps. First, text is extracted from digital objects to produce machine-readable representations of collection materials. Second, the extracted text is used to generate candidate metadata fields corresponding to the descriptive structures used in digital library systems. Third, named entities, such as persons, places, events, and communities, are extracted from the text in order to derive relational signals that extend beyond the item-level record.

This experimental design builds on our previous research on text extraction and classification (Banerjee et al., 2024a; Jadhav et al., 2025; Banerjee et al., 2024b; Miller and Banerjee, 2025). For entity extraction, we use the machine-readable text produced in the previous stage to identify entities such as persons, locations, events, and communities using named entity recognition (NER). Our experimentation will focus on using open-weight language models such as Llama-3 (Llama Team, 2024) and Mistral (Jiang et al., 2023). Using smaller to medium parameter openly available models allows controlled experimentation with direct control over model parameters and inference settings. Prompted examples will be used to guide the models in identifying and extracting entities from the collection materials. We will systematically vary prompts, configurations, and thresholds in order to evaluate metadata generation accuracy, consistency, and efficiency across the collections.

Task 3: Scaling and Evaluation. To evaluate Phase 1, the experimental workflow is applied across multiple collections in the DLP in order to observe how metadata generation behaves across heterogeneous materials. Evaluation focuses on three dimensions: accuracy, consistency, and traceability. For metadata fields that have existing human-authored values, AI-generated metadata will be compared against the human labels to measure agreement and identify conflicts. For authority-controlled fields such as subject headings, we examine whether generated metadata aligns with controlled vocabularies and established descriptive practice. This process allows us to judge the feasibility of AI-generated metadata and find out the limitations for integration when scaled to a production environment.

Consistency is evaluated by examining variation in generated metadata across formats and collection types. Error analysis will be conducted on selected metadata fields to identify systematic generation failures. We will perform error analysis of metadata fields and test for consistency in the generated language across formats and types. We also compute semantic similarity, such as using cosine similarity, BERTScore to quantify agreement across repeated generations of AI-generated metadata.

Traceability is evaluated by examining whether generated metadata can be linked back to identifiable source objects within the collection. This analysis determines whether derived metadata maintains provenance relationships necessary for archival transparency and interpretability. The results of this phase will identify which categories of AI-generated metadata remain sufficiently stable for persistent storage and which require human review or validation. These findings will be documented in a technical guidance manual describing the conditions under which machine-generated metadata can be integrated into operational digital library systems.

Phase 2

Phase 2 investigates Research Question 2 by examining the structural relationships that emerge when machine-generated metadata is applied across multiple collections that can inform how digital library systems organize and connect materials. The collections analyzed in this phase document Appalachian communities in Southwest Virginia and include oral histories, community archives, regional organizational records, and other cultural heritage materials described earlier in this proposal. Using the metadata and extracted entities generated in Phase 1, we analyze how persons, places, events, and communities link materials across otherwise separate collections. This phase therefore evaluates whether large-scale machine-generated metadata can reveal cross-collection relationships that were previously difficult to represent within traditional item-level description.

Task 4: Build Structural Representations. Using the corpus-level entities identified in Phase 1, we extract relationships between entities and represent these relationships using a graph-based structure. Entities such as persons, locations, organizations, and events form the nodes of the representation, while inferred relationships between them form the connecting edges. Relationship extraction will be performed using the LLMs described in Phase 1, with prompting strategies such as zero-shot and few-shot instructions used to identify relationships expressed in the collection materials. The resulting entity-relationship structures provide a representation of how entities interact across documents and collections.

Task 5: Analyze Cross-Collection Entities. Once the structural representation has been constructed, we analyze how entities connect materials across collections. This analysis measures the frequency of shared entities, identifies duplicate or overlapping entity references, and traces connections between materials that originate from different collections. Selected entities such as a historically significant person or a geographic location will be examined in greater detail to understand how relationships propagate across the collections and how these relationships reorganize materials that were previously described only at the item level.

Task 6: Event-Temporal Linkage Analysis. In this task, we focus on temporal and event-centric entities. We test whether AI-generated entities and structural relationships were able to link materials across collections. We direct our efforts to identifying the relationships that have become more apparent after applying entity extraction and relationship linking via knowledge graphs. By applying temporal normalization, a technique that aligns and standardizes data collected over time across various formats to a common, consistent baseline. For example, text normalization can help convert expressions like tomorrow, last year to specific dates. We will examine the impact of temporal normalization on event extraction from distributed episodes. By the end of this task, we will have all events clearly identified and mapped by their order of occurrence. We will share our findings on any event-centric structural mapping as a deliverable from the efforts of this task.

Phase 3

Phase 3 investigates Research Question 3 by examining the entity relationships identifies in Phase 2 enable new ways of engaging with historical documents and cultural memory. As AI-generated metadata

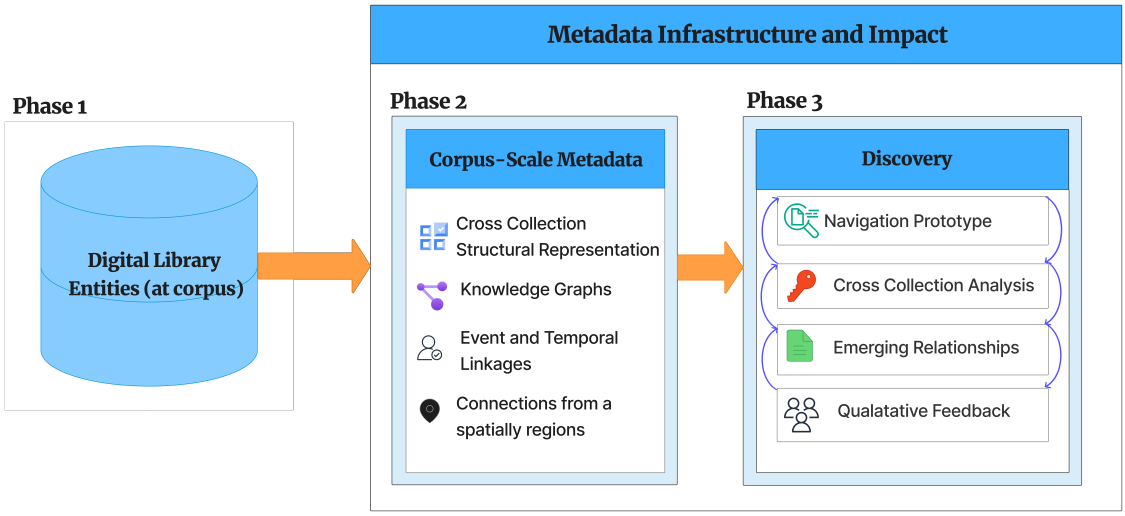

Figure 1: AI-generated metadata as an infrastructural layer, organized into three phases.

entities and relationships become more embedded within the structure of the digital library, new structural relationships emerge that extend beyond traditional item-level metadata. This final phase examines how entity-based relationships support new forms of cross-collection interaction, enabling users to explore materials by connecting people, places, events, and communities.

Task 7: Develop Navigation Prototype. In this task, we develop an experimental prototype implemented within the DLP that demonstrates how entity-based relationships can support navigation across collections. The prototype will allow users to browse shared entities across collections, follow links between related historical events, and traverse connections between materials documenting the same communities. The protype will allow assessment of whether shared entities can link materials across formats (such as text, images, and video). The results of this task will be documented through technical specifications describing the framework for entity-based metadata, along with a field guidance manual and implementation guidance for practitioners working with digital library systems.

Task 8: Cross-collection Discovery Analysis. In this task, we conduct a systematic analysis to determine whether entity-based metadata enables researchers to navigate between archival materials through shared contextual relationships. Using the structural relationships and entity linkages generated in earlier phases, we evaluate whether these connections reveal associations that were that were previously implicit within the collections. The anaysis seeks to expose cross-collection and cross-format pathways that are correlated by entities, thereby reducing the siloed isolation of collections and information within them.

Task 9: Feasibility Assessment of Generated Narratives. Using the navigation prototype and the structural framework, this task examines the feasibility of constructing cross-collection narratives. The analysis examines whether enriched structural metadata supports traceable, evidence-grounded cross-collection interpretation. Testing will include iterative refinements of the prototype informed by discovery analysis and qualitative feedback gathered through focus group sessions. The results of this task will document how cross-collection narrative pathways emerge from the structural metadata and will be summarized in a pathway evaluation report describing the interpretive usability of the system.

Engaging with Stakeholders and Assessing Impact

Throughout the three-year applied research project, we allocate time in each phase to communicate with and involve key stakeholder groups. To maximize the practical relevance of the work, we will engage

with our (1) local metadata experts to refine the machine-generated structural metadata layer, (2) the University Library Deans' Advisory Council to obtain strategic guidance and institutional feedback, and (3) community partners who interact with the digital library collections. In addition to the above groups, we will engage library practitioner and academic research communities through presentations and workshops at venues such as the Association of Southeastern Research Libraries and Code4Lib, as well as through online and in-person workshops hosted at Virginia Tech. Alongside the quantitative evaluation conducted during each phase of the project, we will collect qualitative feedback from our stakeholders participating in our workshops and consultations at key project milestones. This feedback will inform iterative improvements to the workflows and technical approaches developed in the project. Sharing progress throughout the project lifecycle will foster ongoing discussions about the integration of AI in digital library systems and services. Details on the project publication and other dissemination activities are provided in the Project Results section of this narrative.

Project Team

This project team is well positioned to carry out this work, and it brings together expertise in digital libraries, information science, AI, and project management. PI Banerjee and Co-PI Ingram have extensive experience with machine learning and AI applications in digital libraries, and are both affiliated with the Center for Digital Research and Scholarship (CDRS, 2024) at the Virginia Tech University Library, where Co-PI Ingram serves as its director. Through CDRS, the project will have access to interdisciplinary collaborators and research expertise relevant to AI-driven digital library systems. The project will also draw on the university's Advanced Research Computing cluster, which provides high-performance computing resources including sophisticated GPU-enabled systems. In addition, local GPU-equipped servers within the Libraries will support model development, experimentation, and implementation throughout the project. PI Banerjee and Co-PI Ingram will provide methodological and research leadership, while Co-PI Goyne will oversee operational coordination and implementation within the Libraries' digital library environment.

Dr. Bipasha Banerjee will serve as the Principal Investigator and Project Director. Dr. Banerjee is the AI Research Scientist at the University Libraries at Virginia Tech. She will lead the design and implementation of AI and LLM experimentation. Her research lies at the intersection of natural language processing, machine learning, information retrieval, AI and digital libraries. She got her Ph.D. in Computer Science at Virginia Tech in 2024 and is a trained library faculty member with a well-established background in computer science and digital libraries. She will also oversee the progress and ensure that the team adheres to the milestones outlined in the timeline. She will oversee project budget and finances, and also make sure that the project adheres to ethical and privacy standards. She will supervise the GRA for the project, leveraging her experience in mentoring graduate students and fostering a learning environment for research and development.

Jennifer L. Goyne will serve as the Co-Principal Investigator for the project. Jennifer will provide leadership and operational oversight for the project. In this role, Jennifer will not only be responsible for ensuring the successful execution of the project activities and deliverables but also for contributing to and working closely with the principal investigator on the direction, design, and impact of the work. Jennifer is the Assistant Director of Digital Library Development, and she collaborates closely with teams within the Virginia Tech University Library as well as colleagues across the institution. She will leverage her administrative and scholarly leadership in this work to ensure that the impact and success of this project can extend beyond the defined grant timeline.

William A. Ingram will serve as the Co-Principal Investigator for the project. Ingram serves as Associate Dean and Executive Director for Information Technology and Associate Professor at the Virginia Tech University Libraries and as Director of the Center for Digital Research and Scholarship. His research

examines the application of artificial intelligence, natural language processing, and large-scale information retrieval to digital library systems and scholarly corpora. In this project, Ingram will oversee research design and conceptual framing and will guide the development of evaluation frameworks for assessing AI-generated metadata at corpus scale. He will ensure that project workflows and outputs remain auditable, reproducible, and reusable, enabling the project's methods and findings to be applied across digital library platforms.

Graduate Research Assistant A Computer Science Ph.D. student at Virginia Tech with an interest in pursuing machine learning and AI research will be supported through this project. They will undertake tasks such as data selection, organizing and preprocessing, developing AI-integrated experimentation prototypes, and implementation of code, models and algorithms. The GRA will maintain detailed records of methodologies, experimental designs, and results, and contribute to project reporting, and dissemination activities, including article preparation, conference presentations, and workshops. The student will participate fully in weekly project meetings and collaborative research activities. The GRA benefits from expert mentorship which will be structured around the NSF plans (NSF, 2024), including regular advising meetings, training opportunities in responsible and reproducible research practices, interdisciplinary collaboration, and opportunities for professional development. By working in the library, the student will gain unique interdisciplinary experience that spans across digital libraries, artificial intelligence, natural language processing, and research services.

PROJECT RESULTS

The intended results of this project are a set of practical resources that enable libraries to evaluate and adopt AI-assisted approaches for working with large digital collections. These results include a practitioner-oriented manual describing methods for integrating LLMs into digital library workflows, open-source software and documentation that demonstrate how these methods can be implemented, as well as workshops and presentations designed to introduce these approaches to library professionals and researchers.

By documenting approaches for extracting entities and generating structural representations of digital collections using large language models, the project provides libraries with practical guidance for extending existing digital library services to support large-scale computational analysis. These methods allow libraries to experiment with AI-assisted metadata enrichment and structural analysis while maintaining transparency about how these techniques are applied.

Dissemination will occur through multiple channels aimed at both practitioners and researchers. Project findings will be shared through workshops and presentations within the library and digital scholarship communities, including venues such as Code for Librarians (Code4Lib) and the Association of Southeastern Research Libraries (ASERL). The research results will also be disseminated through peer-reviewed publications and conference presentations in venues such as ACM/IEEE Joint Conference on Digital Libraries (JCDL), Theory and Practice of Digital Libraries (TPDL), ACM Special Interest Group on Information Retrieval (SIGIR), IEEE BigData, and Journal of the Association for Information Science and Technology (JASIST). In addition, manuscripts and project outputs will be deposited in open repositories, including VTechWorks, and arXiv, ensuring broad public access.

We aim is to disseminate information widely and regularly throughout the project lifecycle. Our deliverables include the following:

Resources for library practitioners: We are committed to advancing the training and professional development of the library workforce. In line with this commitment, we will host workshops at our university to gain qualitative feedback and train library professionals.

1. **Comprehensive Practitioner's Manual** We will create and share a practitioner's manual regarding using and integrating LLMs into library services to extract entities and create a structural

representation layer within digital libraries. This manual will contain detailed information about our approaches, methodology, experiments and results. Predominantly developed with library practitioners in mind, this guide is also intended to address the needs of the digital library.

2. **Outreach and networking** We plan to engage more in library events, such as speaking at the ASERL professional development series. This is particularly important, as our collection comprises rural Appalachian communities and would therefore be relevant to the Southeastern research library community.
3. **Presentations and Workshops** We also plan to host workshops at libraries, archives, and technologist communities such as Code4Lib to demonstrate how our work can be adopted in practice by other digital libraries. Our goal through these aforementioned opportunities is to promote the usage and integration of LLMs into library services.
4. **Open-Source Software and Documentation** All software and coding notebooks will be openly available and will be accompanied by detailed documentation to reproduce the results from our project. This also includes all software prototypes to help other digital libraries implement our approach to their collections, therefore testing and measuring its effectiveness.

Sharing with academic research community: We plan to engage in publishing through conference proceedings and high-impact journals, with prioritizing open-access journals. We plan to submit our findings to venues such as JCDL, TPDL, SIGIR, IEEE BigData, and JASIST. We also want to present and engage with communities such as the Advanced Information Collaboratory (AICollaboratory, 2020), a group encompassing information and library scientists and archivists to help disseminate our findings to academic practitioners. We will also deposit the accepted version of the manuscripts to the IMLS-directed repository, our institutional repository (VTechWorks), and arXiv to support the agency's public access guidance.

Graduate Student Mentorship: A significant amount of the budget is allocated to supporting a GRA. While the GRA will help meet project goals by performing the tasks outlined in the grant, it also offers the student's research and professional career by participating in this interdisciplinary research. This will help the student advance their own dissertation agenda, gain an educational degree, as well as valuable research experience. Participating in publishing also helps the student gain research visibility and helps the project demonstrate its impact.

The project is designed so that its benefits extend beyond the grant period. By releasing all software, documentation, and methodological guidance as openly accessible resources, the project creates reusable infrastructure that other institutions can adopt or extend. Continued engagement with professional communities through workshops, conference presentations, and practitioner networks will further support the ongoing dissemination and refinement of these approaches. In this way, the project establishes a foundation for future work in applying large language models to digital library collections while enabling other institutions to build upon the methods and resources developed through the project.

Overall, our project aims to sustain benefits beyond the scheduled completion by opening all our research materials for reuse by other institutions and digital libraries. We are committed to serve our community and the public and adhere strictly to all public access and ethical guidance. We will also continue participating in events and workshops across the library and information science community to help convey our findings.

Library-wide, Machine-Generated Metadata as First-Class Structural Components

Bipasha Banerjee, Jennifer L. Goyne & William A. Ingram

Schedule of Completion

	Year 1 (2026–2027)											
Activities and Milestones	**S**	**O**	**N**	**D**	**J**	**F**	**M**	**A**	**M**	**J**	**J**	**A**
Administrative Tasks												
Launch Project *Lead: Banerjee, Goyne and Ingram*												
Hire Graduate Research Assistant *Lead: Banerjee & Goyne*												
Review Public Access Plan *Lead: Banerjee*												
Develop And Review Project Work Plan *Lead: Banerjee*												
Review Monthly Budget Reports *Lead: Banerjee, Goyne*												
Create And Update Project Website/Documentation *Lead: GRA*												
Annual reporting *Lead: Banerjee, Goyne and Ingram*												
Phase 1 — Investigating RQ1												
Data Selection *Lead: Banerjee and Goyne*												
Experimental Workflow *Lead: Banerjee*												
Scaling and Evaluation *Lead: Goyne and Banerjee*												
Document Findings *Lead: Banerjee, Goyne and Ingram*												
Phase 2 — Investigating RQ2												
Build Structural Representations *Lead: Banerjee, Goyne*												
Reporting and Sharing												
Draft and Submit Conference Poster/Short Paper *Lead: Banerjee, Goyne and Ingram*												
Draft and Publish Guidance Manuals *Lead: Banerjee, Goyne and Ingram*												

	Year 2 (2027–2028)											
Activities and Milestones	**S**	**O**	**N**	**D**	**J**	**F**	**M**	**A**	**M**	**J**	**J**	**A**
Administrative Tasks												
Review Public Access Plan *Lead: Banerjee*												
Review And Update Project Work Plan *Lead: Banerjee*												
Review Monthly Budget Reports *Lead: Banerjee, Goyne*												
Update Project Website/Documentation *Lead: GRA*												
Annual Reporting *Lead: Banerjee, Goyne And Ingram*												
Phase 2 — Investigating RQ2												
Build Structural Representations *Lead: Banerjee, Goyne*												
Analyze Cross-collection Entities *Lead: Banerjee*												
Event-Temporal Linkage Analysis *Lead:Goyne*												
Reporting and Sharing												
Draft and Submit Conference/Journal Paper *Lead: Banerjee, Goyne and Ingram*												
Draft and Submit Conference Workshop Proposal *Lead: Banerjee, Goyne and Ingram*												
Draft and Publish Guidance Manuals *Lead: Banerjee, Goyne and Ingram*												

	Year 3 (2028–2029)											
Activities and Milestones	**S**	**O**	**N**	**D**	**J**	**F**	**M**	**A**	**M**	**J**	**J**	**A**
Administrative Tasks												
Review Public Access Plan *Lead: Banerjee*												
Review And Update Project Work Plan *Lead: Banerjee, Goyne*												
Review Monthly Budget Reports *Lead: Banerjee, Goyne*												
Update Project Website/Documentation *Lead: GRA*												
Final Reporting *Lead: Banerjee, Goyne And Ingram*												
Project Wrap Up *Lead: Banerjee*												
Phase 3 — Investigating RQ3												
Develop a Navigation Prototype *Lead: Banerjee*												
Cross-collection Discovery Analysis *Lead: Goyne*												
Feasibility Assessment of Generated Narratives *Lead: Banerjee and Goyne*												
Reporting and Sharing												
Draft and Submit Conference/Journal Paper *Lead: Banerjee, Goyne and Ingram*												
Draft and Submit Workshop Proposal *Lead: Banerjee, Goyne and Ingram*												
Conduct Workshop *Lead: Banerjee, Goyne and Ingram*												