



NATIONAL LEADERSHIP GRANTS FOR LIBRARIES

Sample Application LG-259770-OLS-26
Project Category: Applied Research

Virginia Polytechnic Institute and State University Libraries

Amount awarded by IMLS: \$613,106
Amount of cost share: \$0

The Virginia Polytechnic Institute and State University Libraries will collaborate with the Virginia Tech Sanghani Center for AI and Data Analytics, and Northwestern University Libraries to research the challenge of inconsistent descriptive metadata across digitized collections and the risks to privacy and institutional control inherent in proprietary artificial intelligence models. This research will investigate how agentic artificial intelligence architectures combining structured reasoning with tool-based verification can be designed to meet library standards for transparency and local stewardship, producing a validated design framework for privacy-preserving artificial intelligence, a benchmarking methodology for evidence quality, and an open-source toolkit for nationwide institutional adaptation. The project will produce a documented reference architecture that separates curatorial policy from model choice, allowing institutions to adopt improved artificial models over time without redesigning their metadata workflows. The project will develop an open library of agentic skills (prompt templates, workflow sequences, and grounding configurations) designed to process complex archival materials and reduce factual errors in metadata creation.

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.

Sovereign Library Intelligence (SLI): An Agentic AI Framework for Trustworthy Library Collections Stewardship and Access

1 Introduction

Virginia Tech University Libraries, in collaboration with the Virginia Tech Sanghani Center for AI and Data Analytics and Northwestern University Libraries, requests \$613,106 in funding from the Institute of Museum and Library Services (IMLS) National Leadership Grants for Libraries (NLG-L) program for a three-year applied research project addressing the persistent challenge of inconsistent descriptive metadata across digitized collections and the risks to privacy and institutional control inherent in proprietary AI models. This research investigates how agentic AI architectures combining structured reasoning with tool-based verification can be designed to meet library standards for transparency and local stewardship, producing a validated design framework for privacy-preserving AI, a benchmarking methodology for evidence quality, and an open-source toolkit for nationwide institutional adaptation. The investigation is guided by four central research questions (RQ): **(RQ1)** which combinations of structured reasoning steps and verification methods most effectively reduce factual errors under real library constraints; **(RQ2)** how AI-generated descriptions of text, images, and video can be directly linked to the specific portions of archival materials that support them; **(RQ3)** To what extent can an expert-validated toolkit accurately quantify the groundedness and fidelity of AI-generated archival metadata; and **(RQ4)** under what conditions a locally hosted, domain-focused model demonstrates higher factual precision and institutional alignment compared to representative proprietary baselines.

2 Project Justification

2.1 Alignment with Agency and Program Goals

This applied research project directly supports IMLS Agency-Level Goal 3: Advance Collections Stewardship and Access, specifically aligning with Objective 3.2: Promote access to museum and library collections. Concurrently, the project fulfills the NLG-L Program Goal to develop replicable practices, models, or tools to strengthen library and archival services, addressing Objective 1.3: Provide broad access to and preservation of information and collections through libraries and archives. By investigating verifiable AI-assisted metadata workflows, this research enables libraries to expand digital access to historical collections while explicitly addressing priorities regarding artificial intelligence, privacy, and collections stewardship. Furthermore, as the nation approaches its 250th anniversary, this research responds to the strategic priority to promote libraries as they teach Americans about their heritage by ensuring digital access is anchored in factual accuracy and verifiable provenance.

2.2 The Need, Challenge, and Opportunity

The library and archival fields face a significant and practical challenge: unlocking millions of poorly described digital assets (“dark data”) [1] to facilitate meaningful public discovery, without compromising professional standards of truth and stewardship. While libraries are increasingly turning to Artificial Intelligence (AI) to automate metadata generation, an environmental scan of 2025–2026 scholarship reveals a critical “trust gap” in the field.

Current research identifies that standard generative models are highly prone to hallucinations, including fabricated citations and factual inconsistencies [2,3]. While Retrieval-Augmented Generation (RAG) systems aim to mitigate this issue [4], early implementations still struggle with nuanced interpretation and conflicting evidence in complex domains [5]. Librarians are increasingly burdened with the massive cognitive load of verifying AI-generated outputs that contain plausible but entirely imaginary sources, creating a need for task-specific, reliable systems rather than general-purpose text generators.

This technical limitation is compounded by an ethical and infrastructural threat: the reliance on proprietary, cloud-based architectures. The current paradigm of commercial AI often treats cultural heritage data as training material, ingesting institutional metadata into irreversible models without consent. Scholars identify this dynamic as a risk of institutional data extraction, where the control of digital infrastructure and intellectual property is surrendered to proprietary, third-party AI corporations [6]. This dynamic is a form of algorithmic dependence where the digital infrastructure and local cultural

heritage are harvested to reinforce unequal power structures [7]. Institutions risk trading long-term control of their cultural heritage for short-term automated convenience.

This critical need presents a vital opportunity. By researching how locally hosted, multi-agent AI systems can interpret archival materials through traceable verification, this project provides the empirical foundation needed for libraries to transition from passive consumers of proprietary AI to sovereign stewards of their own digital intelligence.

2.3 Differentiation, Complementarity, and Scholarly Context

This project establishes the benchmarking methodologies and design frameworks required to transition AI applications in the libraries and archives sector from experimental pilots to standardized professional practice. The SLI framework introduces three core technical innovations: (1) **agentic verification pipelines** that mimic professional archival reasoning; (2) **evidence-linked metadata generation** where each field is tied to specific source segments; and (3) **institutionally sovereign AI deployment** enabling local model hosting without reliance on commercial cloud services. This research builds upon and extends current scholarship and field-wide initiatives in the following areas:

- **Extending AI Ethics Research into Technical Architecture:** This study utilizes the foundational research of the IMLS-funded *Responsible AI in Libraries and Archives* (LG-252307-OLS-22) project. While that initiative established ethical frameworks and harms analysis tools, our research investigates the technical architectures required to implement those values in a library setting. We specifically research the local deployment of open-weights models and verifiable reasoning steps to ensure ethical alignment without third-party data ingestion.
- **Responding to Field-wide Calls for Trustworthy Metadata Workflows:** This project responds to the urgent mandate for the Libraries, Archives, and Museums (LAMs) community to address the impact of generative AI on collection integrity [8]. By engineering traceable verification tools, we fulfill the national call for “collective strategies centered upon human judgment and collaboration” [8] rather than relying on a single industry standard. This technical approach directly addresses the reasoning limitations identified in current generative workflows by shifting the AI from a passive discovery tool to a proactive, collaborative processor of archival evidence.
- **Applying Emerging AI Governance Policies at the Infrastructure Level:** This research builds on emerging institutional responses to commercial data extraction. The University of Virginia’s Archival AI Protocol, published in 2026 [9], established an early systematic framework for research libraries to evaluate and respond to AI companies seeking access to archival collections for model training. The Protocol has been adopted or reviewed by more than 40 institutions, demonstrating that the field is actively seeking governance tools to assert institutional control over collections data. The SLI project directly complements this policy-level work by engineering the localized technical infrastructure required to enable sovereign AI adoption at the operational level.
- **Applying provenance standards to metadata:** The research adapts provenance and authenticity frameworks discussed in [8], including initiatives from LC Labs and the Coalition for Content Provenance and Authenticity (C2PA), to verify AI-generated metadata. While existing initiatives within the library and archival community explore C2PA for digital asset authenticity, this project applies these standards to AI-generated metadata. We investigate how “Content Credentials” can link automated descriptions directly to the specific archival evidence that documents our shared national heritage. This ensures that the metadata used to interpret the defining chapters of the American story remains verifiable, fact-based, and preserved for a nation of citizens seeking to engage with their founding history.
- **Advancing from linear to agentic RAG:** This study builds upon the technical implementation of linear Retrieval-Augmented Generation (RAG) for collection discovery achieved in project (LG-256703-OLS-24). While that initiative successfully deployed RAG, this research investigates the next technical iteration: agentic AI. Traditional RAG systems often fail to reconcile conflicting records within historical sources or verify citations against primary evidence, providing single, synthesized answers that obscure contradictions within primary historical sources, a limitation that the SLI framework addresses through multi-agent verification cycles [10, 11]. This project moves beyond simple retrieval by using specialized agents to resolve contradictions in primary sources and systematically verify citations against primary evidence. This approach directly addresses the reasoning limitations identified in standard generative workflows by shifting the AI from a passive discovery tool to a proactive, collaborative processor of archival evidence.

- **Implementing sovereign AI models:** The project applies the global sovereign AI paradigm to the constraints of the library field. Current research emphasizes the need for public institutions to maintain control over technical infrastructure and data to avoid vendor dependency and “digital colonialism” [12]. This study evaluates the performance of open-weights models on local institutional hardware compared to proprietary cloud baselines, establishing benchmarks for performance, privacy, and long-term sustainability.

2.4 Project Beneficiaries

The research results, including a validated design framework, an evidence-quality benchmarking methodology, and an open-source toolkit, will provide sustained benefits to three primary groups:

1. **The American Public and Researchers:** By improving the traceability and reliability of automated archival descriptions, the public gains more equitable, accurate, and meaningful access to shared historical collections. Direct evidence-linking protects users from AI-generated misinformation, reinforcing the library’s role as a trusted civic center of learning.
2. **Library and Archival Professionals:** Metadata librarians and archivists will benefit from research into explainable AI interfaces. By determining what level of explanation best supports professional review, the project reduces the time and cognitive load required to verify automated metadata, ensuring AI operates as an assistant rather than a replacement for human judgment.
3. **The National Cultural Heritage Sector:** The project democratizes access to advanced technology by prioritizing approaches that can be deployed locally using modest computing resources. By producing an open-source toolkit, institutions of all sizes will gain a replicable, empirically validated blueprint to implement trustworthy AI without relying on expensive, opaque commercial vendors.

3 Project Work Plan

3.1 Theoretical Framework and Conceptual Foundations

To ensure generalizability and methodological rigor, the SLI project builds on prior scholarship in AI, archival practice, and information governance and is anchored in three intersecting domains: Agentic Artificial Intelligence Theory, Expert Workflow Analysis, and Institutional Sovereignty.

Agentic Artificial Intelligence Theory

The project fundamentally differentiates itself from previous library AI initiatives by utilizing Agentic AI rather than reactive text-generation models. While traditional LLMs operate on a transactional basis, Agentic AI systems are engineered for automated task-execution and multi-step coordination. This design is realized through the Archival Agentic Assistant (AAA), which utilizes three coordinated agents: a **Planner Agent** to decompose archival processing tasks, an **Extraction Agent** to identify metadata entities, and a **Verification Agent** to validate extracted information against authority files. This architecture achieves automated workflow orchestration, persistent context memory, and verifiable task execution grounded in explicit archival evidence [13–16].

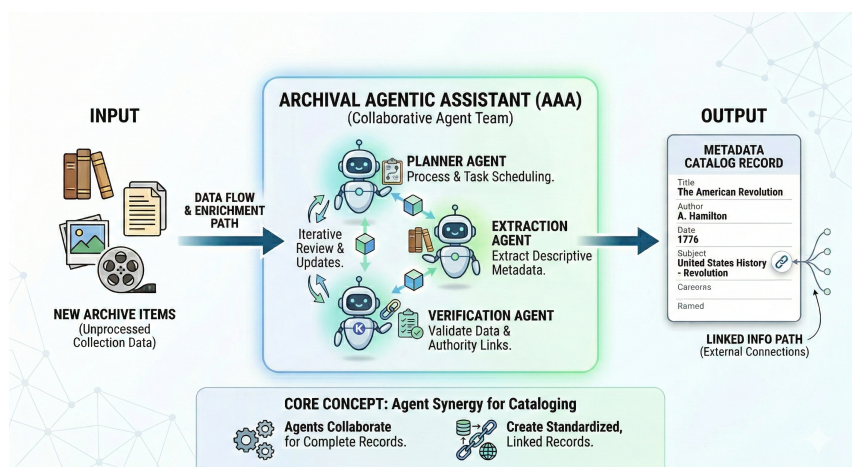


Figure 1: Archival Agentic Assistant (AAA) Collaborative Logic.

Drawing upon both the symbolic lineage of algorithmic decision-making and the neural lineage of statistical learning, the SLI framework evaluates distinct multi-agent control structures. To mitigate escalating computational costs and prohibitive hardware requirements, the framework enforces bounded, closed-loop orchestration with explicit stopping criteria and controlled token allocation. Within this framework, distinct algorithmic personas operate collaboratively using a Directed Acyclic Graph (DAG) task-graph. Such multi-agent coordination enables the orchestration of complex workflows [17, 18]. As illustrated in Figure 1, the Archival Agentic Assistant (AAA) is engineered to evaluate conflicting historical evidence by executing non-linear reasoning paths to identify the most probable metadata attributes [19-21]. A planner agent decomposes complex documents for process and task scheduling. An extraction agent isolates descriptive metadata, while a verification agent validates data and authority links against the sovereign knowledge base to prevent hallucinations. By shifting the AI from a passive instructional tool to a proactive co-processor of archival evidence, the project mirrors the complex, multi-step verification processes inherent in professional archival description.

Expert Workflow Analysis

To accurately design the workflows of these autonomous agents, the project will analyze the existing workflow processes and documentation from the project team and generate the work needed to inform system design. This analysis represents a core component of the research, applied to understand complex work processes, inform the design of decision support systems, and elicit the procedural knowledge that subject matter experts use to perform high-level tasks. Archival description, like clinical diagnostic reasoning, operates under cognitive constraints, including limitations in working memory [22] and vulnerability to premature closure [23]. A systematic analysis of current workflows provides the mechanism to deconstruct how an expert archivist overcomes these limitations when approaching an unfamiliar historical document.

The project will utilize established protocols for documentation analysis and workflow verification. By analyzing the professional practices of metadata librarians as recorded in project team documentation, the research team will isolate specific strategies, such as "relevance screening" (reducing the amount of information to sort through by making relevant cues salient) and "minimizing memory load" (tracking the status of tasks through structured division of labor). These expert strategies will then be computationally mapped onto the AI agents. For instance, if the analysis of existing workflows reveals that expert archivists invariably verify geographic locations against an established authority file before assigning a subject heading, the research will test whether encoding this procedural check as a programmatic constraint in the Verification Agent produces measurably more accurate metadata outputs compared to unconstrained generation. By translating documented human strategies into machine logic, the SLI framework embeds expert methodologies directly into the machine-learning pipeline, ensuring that the AI's reasoning aligns with established library science paradigms and addresses the nuanced complexities of archival metadata.

Institutional Sovereignty and Governance

The third theoretical pillar guiding this research is Institutional Sovereignty. Libraries and archives that collect and preserve human knowledge are essential infrastructures that must strategically adopt emerging technologies, including artificial intelligence, to advance academic practice. However, they are increasingly at risk when commercial AI tools are adopted without adequate evaluation, governance, and transparency. Commercial AI subscriptions create unequal access and compliance risks through opaque processing algorithms and non-transparent data hosting. When cultural heritage institutions feed archival documents into proprietary cloud models, they risk exposing sensitive information and inadvertently contributing their curated intellectual property to external corporate training datasets without compensation or consent.

The concept of knowledge sovereignty suggests that institutions must retain meaningful control over their digital assets and computational infrastructures. To explore how this principle may be operationalized in AI-enabled archival systems, the SLI project investigates the use of locally hosted, open-weight foundation models, such as Llama 3 [24], operating within institutional server environments. This study conceptualizes a "gateway pattern" of AI governance, in which all model interactions are, by default, mediated through institutional infrastructure. Rather than treating governance as a post-hoc audit of outputs, the project examines how algorithmic constraints can be embedded at the tool and interface level. It further analyzes the security implications of agentic architectures, including risks such as prompt injection from uncured archival text [25] and unintended tool execution. The framework evaluates a defense-in-depth model incorporating API restrictions, schema validation, provenance controls, and audit logging [26] as mechanisms for institutional oversight. Finally, the project investigates transparency not as rhetorical openness but as traceable evidentiary linkage, surfacing textual support and historical

conflict rather than collapsing archival ambiguity into synthetic certainty.

3.2 Methodology

The methodological approach combines computational systems engineering, natural language processing (NLP) benchmarking, and systematic algorithmic auditing. The study is designed to ensure that the findings are robust, replicable, and highly generalizable across different cultural heritage institutions.

Experimental Testbed and Data Sources

To empirically validate the SLI framework, the project establishes a multi-tiered testbed prioritizing founding-era documents from the National Archives and the Digital Public Library of America (DPLA). By linking AI-generated metadata directly to evidence within these records, the project fulfills the IMLS strategic priority to teach Americans about their heritage and history during the Semiquincentennial. The data acquisition strategy progresses from high-volume baseline establishment to localized stress-testing:

1. **Open National Collections:** By ingesting high-volume public domain collections, specifically records from the DPLA, Chronicling America, and founding-era documents from the National Archives, the project establishes a reproducible baseline with sufficient data volume for initial model alignment while ensuring findings remain reproducible by peer institutions.
2. **Partner Institutional Collections:** To increase data complexity, the framework will process specialized datasets sampled from partner institutions. Virginia Tech contributes localized archives, including uncatalogued Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) texts detailing Appalachian history and early American frontier industrialization. Northwestern University provides complementary collections documenting transportation history and global civic movements. These repositories introduce ambiguity and evidentiary variation necessary to calibrate the agents' extraction and verification protocols for mid-sized and regional archives.

Beyond textual sources, the dataset includes multimedia materials, such as transcribed oral histories and descriptive metadata for historical photographs, to evaluate multimodal processing across formats [27].

Multi-Agent System Architecture and Provenance Linking

This phase involves the development and rigorous evaluation of the SLI framework's technical architecture. The research will investigate a multi-agent orchestration model designed to mirror the expert workflows identified during the Phase 1 workflow analysis.

Sovereign Knowledge Base Integration

To ensure the framework's Verification Agent grounds its reasoning in established archival standards, the project will ingest authoritative ontologies into a locally hosted vector database. This "Sovereign Knowledge Base" will be explicitly seeded with the Library of Congress Subject Headings (LCSH) [28], the Faceted Application of Subject Terminology (FAST), the Getty Art & Architecture Thesaurus (AAT), and the partner institutions' localized finding aids. This guarantees the autonomous agents reconcile extracted entities against standardized professional vocabularies rather than relying on the generalized, often inaccurate semantic approximations inherent to commercial LLMs.

The technical orchestration of the framework is visualized in Figure 2. This architecture enforces a gateway-based AI governance pattern by keeping all processing within the institution's secure perimeter. Archival materials are ingested into a bounded closed-loop environment where agents interact with a sovereign knowledge base containing authority files and

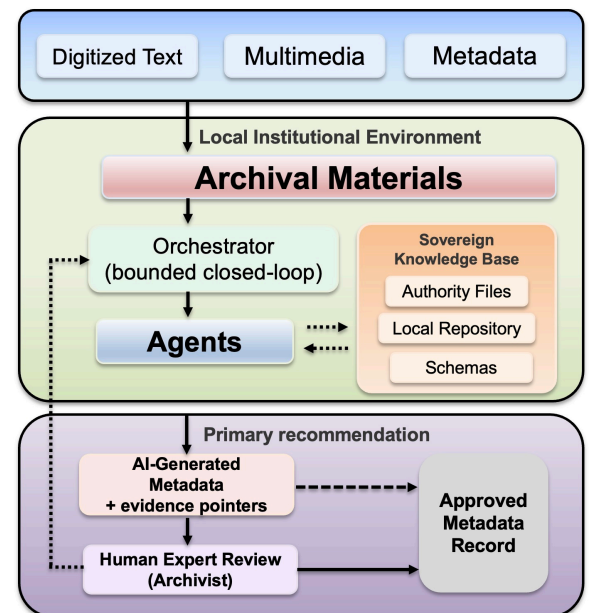


Figure 2: SLI Multi-Agent Architecture.

local repository data. This process results in AI-generated metadata and evidence pointers. These outputs are presented for human expert review to ensure archival rigor and factual accuracy before final cataloging [29].

Experimental Agent Roles

Utilizing this integrated knowledge base, the project will test a sequential processing pipeline to determine the most effective configuration for archival discovery, consisting of the following experimental roles:

- **Extraction Agent Evaluation:** This agent will be configured to identify named entities, dates, and locations within OCR/HTR outputs. The project will assess the accuracy of these extractions when guided by the bespoke algorithmic instructions derived from Phase 1 findings.
- **Adversarial Verification Testing:** To address **RQ1**, a Verification Agent will be implemented to cross-reference extractions against the sovereign knowledge base. This stage of the research aims to quantify the extent to which strict, closed-domain cross-referencing can mitigate or eliminate algorithmic hallucinations in complex archival contexts. Success requires a statistically significant reduction in hallucination rates compared to standard RAG baselines.
- **Provenance Linking Feasibility:** To address **RQ2**, the project will evaluate the Verification Agent’s ability to generate precise provenance links, such as bounding-box coordinates, for extracted entities. The research will determine if this method effectively links generated metadata back to its source in the digital surrogate, providing a verifiable and transparent audit trail.

Analytical Evaluation and Fidelity Quantification

Addressing **RQ3** requires the development of a novel, advanced analytical toolkit to measure AI performance. In long-horizon, multi-step agentic deployments, small reasoning errors compound, and non-determinism can critically undermine reproducibility. Therefore, the methodology moves beyond standard final-answer scoring (e.g., BLEU [30] or ROUGE [31]) to implement trace-based evaluation of intermediate reasoning steps. The computational toolkit will explicitly audit the intermediate reasoning steps of the agents, calculating two domain-specific metrics:

- **Groundedness:** A strict mathematical representation of the extent to which every generated metadata tag or descriptive summary can be traced exclusively to the provided source text [32, 33]. The algorithm heavily penalizes the generation of any unsourced historical claims or external assumptions not present in the original artifact.
- **Fidelity:** The semantic accuracy of the generated description in capturing the true historical and contextual meaning of the document. This is measured by computationally comparing the AI’s output against the existing “gold-standard” metadata schemas (e.g., Dublin Core [34], MARC [35], or Encoded Archival Description [36]) crafted by expert human catalogers. Automated scores must correlate significantly ($r \geq 0.75$) with expert archivist judgment.

Comparative Benchmarking and Multi-Model Routing

To address **RQ4**, the calibrated toolkit will conduct a systematic comparative analysis. The outputs from the locally hosted sovereign agentic framework will be benchmarked against leading general-purpose commercial cloud models current at the time of testing, such as the most recent iterations of the OpenAI GPT and Anthropic Claude series. Importantly, the SLI architectural blueprint is engineered to be model-agnostic. This allows institutions to update underlying cognitive engines by swapping current open-weights models for future and more efficient iterations without requiring a full redesign of the metadata pipeline. Furthermore, the project will evaluate an adaptive multi-model routing and aggregation baseline as an advanced systems-level extension. In this variant, the SLI framework will use a prior confidence score to dynamically decide whether to route a metadata extraction request to a single model or to aggregate responses from multiple expert models using support-set-based aggregator selection. This empirical comparison will provide the library field with the first systematically collected empirical evidence on the performance tradeoffs, computational costs, and operational conditions of locally hosted sovereign infrastructure compared to commercial cloud baselines.

Institutional Research Capacity and Infrastructure

Virginia Tech’s Advanced Research Computing (ARC) unit provides the secure, local High-Performance Computing (HPC) environment to host, fine-tune, and orchestrate the project’s sovereign AI models within the gateway governance pattern. AI research consultation is led through the Sanghani Center for AI and Data Analytics. Both units have committed institutional support to this project (see letters of support, Supportingdoc3). This infrastructure absorbs the computational overhead of Small Language Model (SLM) optimization, so the SLI Toolkit can run on the modest hardware typical of mid-sized and rural libraries. Northwestern University serves as the independent validation site, where professional archivists verify AI-generated metadata against gold-standard records.

Approach to Securing Institutional Review Board (IRB) Approval: The evaluation of AI-generated metadata constitutes minimal-risk human subjects research involving professional archivists at Virginia Tech and Northwestern University. As an initial step, the Principal Investigator has already completed the required human subjects research training. All other personnel engaged in human subjects research activities will complete the required training prior to participation. A cross-institutional protocol will be submitted to the Virginia Tech IRB, which will serve as the lead institution and coordinate reliance agreements across partner sites. No human subjects research activities will begin until IRB approval is granted.

3.3 Project Phases and Tasks

The project will be executed in three sequenced annual phases designed to move from **(Year 1)** a defensible, evaluation-ready testbed and formalized expert reasoning, to **(Year 2)** a locally deployable sovereign multi-agent prototype, and finally to **(Year 3)** validated auditing and comparative benchmarking that answer the research questions and support adoption. The ordering of tasks is intentional: each task produces a concrete output required by the next stage, ensuring the work plan is executable, schedulable, and directly tied to **RQ1–RQ4**.

Year 1: Testbed Curation and Workflow Analysis. Year 1 establishes the experimental foundation and captures the professional logic required to design the system. The project begins by ingesting the baseline national collections and selecting representative partner datasets relevant to the U.S. Semiquincentennial, as defined in the Experimental Testbed, while executing the required data-use and access agreements (**Task 1**). Dataset selection will be guided by explicit criteria needed to answer the research questions, including variation in format and material type (e.g., OCR/HTR text, page images, and existing descriptive records), the availability of gold-standard human-created metadata for evaluation, and any rights or access constraints that require local controls. Once selected, the team will normalize and harmonize the datasets into a shared schema, document transformations, and partition gold-standard records into training, development, and test splits, while maintaining partner-controlled security boundaries for any restricted materials (**Task 2**).

In parallel with testbed preparation, the research team will analyze partner metadata workflows and archival documentation to elicit how professional staff screen relevance, verify claims, manage uncertainty, and apply authority practices in descriptive work (**Task 3**). These findings will then be translated into explicit, implementable logic trees and decision rules that define the constraints under which an agentic pipeline should operate in an archival setting—for example, specifying verification checkpoints and conditions under which the system must defer, surface uncertainty, or require evidence before writing a field (**Task 4**). By the end of Year 1, the project will have produced a secure, normalized testbed with evaluation-ready splits and a versioned specification of professional reasoning logic that directly drives system engineering in Year 2.

Year 2: Multi-Agent Architecture Engineering and Provenance Linking. Year 2 converts the Year 1 reasoning specification into a working computational framework and validates it through partner evaluation cycles. Using the logic trees produced in **Task 4** as binding design constraints, the project team will investigate how a multi-agent orchestration pipeline and retrieval layer impact metadata fidelity by scripting RAG workflows that enable the Planner Agent and Extraction Agent to operate over the sovereign knowledge base and curated testbed rather than unconstrained generation (**Task 5**). The system will then be extended with two capabilities that directly address the project’s research questions. To address **RQ1** (reducing factual errors), the project evaluates the effectiveness of an adversarial Verification Agent that cross-checks extracted claims against the sovereign knowledge base and source evidence, enforcing constraints that prevent unsupported assertions from being written into final metadata outputs (**Task 6**). To address **RQ2**, the project investigates the precision and professional utility of the Verification Agent’s provenance linking capabilities, which link extracted fields to explicit evidence, such as OCR spans and page-image coordinates, to determine the minimum evidence threshold required for professional archival auditability (**Task 7**).

Using the SLI prototype, the research will conduct automated auditing cycles to measure the gap between AI-generated metadata and gold-standard human records, quantifying error categories and omission patterns across the testbed to guide iterative refinement (**Task 8**). The project team will use this structured feedback to iteratively refine prompts, constraints, retrieval settings, and system behaviors in controlled cycles, producing a stable, partner-validated prototype and a reusable library of archival-specific prompts and constraint patterns. By the end of Year 2, the project will deliver a functional, locally deployable prototype and evidence of refinement grounded in partner evaluation.

Because the SLI toolkit is engineered around efficient SLMs [37], its usability must be validated outside of resource-rich research universities. In Year 2, the project will host technical webinars to demonstrate the operational requirements of the locally hosted SLM framework. These sessions will inform practitioners that the toolkit is available for testing within their own environments. By inviting external evaluation, the team will gather professional feedback to help generalize common technical needs and refine the system’s hardware requirements. This ensures the Year 2 outreach provides the data needed for the comparative benchmarking scheduled for Year 3.

Year 3: Evaluation Toolkit Calibration and Comparative Benchmarking. Year 3 develops validated evaluation infrastructure, answers **RQ3** and **RQ4** through reproducible benchmarking, and packages results for broad adoption. The research will develop and validate an Archival AI Auditing Toolkit by testing whether automated algorithms that compute domain-specific Groundedness (the extent to which outputs are supported by explicit evidence) and Fidelity (semantic and structural alignment with gold-standard human descriptions and schemas) scores reliably approximate professional expert judgment (**Task 9**). To answer **RQ3**, an expert review panel will manually assess a sample of system outputs; these judgments will be used to calibrate the toolkit’s scoring weights and validate that the automated measures reliably approximate professional evaluation (**Task 10**). With a calibrated toolkit in place, the project will answer **RQ4** by conducting controlled comparative benchmarking, evaluating the locally hosted sovereign framework against representative commercial cloud baselines under matched conditions (same testbed, same tasks, same scoring instrumentation) to quantify performance, reliability, and privacy/sovereignty tradeoffs (**Task 11**).

The project concludes by disseminating the validated design framework, the open-source SLI framework and auditing toolkit, and a comprehensive Practitioner’s Implementation and Governance Manual (**Task 12**). These resources provide a replicable blueprint for national adoption, supported by peer-reviewed publications documenting the framework’s performance and sustainability. In Year 3, the team will use a remote collaboration model with Northwestern University and Year 2 participants to finalize these materials based on real-world deployment findings. Consolidating this feedback ensures that the SLM framework and auditing toolkit remain reproducible and effective for institutions nationwide, regardless of geographic location or local technical capacity.

3.4 Project Evaluation

The SLI project employs a two-tier evaluation strategy to monitor progress and assess research outcomes throughout all three project years. Formative evaluation ensures the work plan remains on track and produces usable intermediate artifacts; summative evaluation measures whether the project’s four research questions have been answered with sufficient evidence for field adoption.

Formative Evaluation. At the close of each project year, the PI and Co-PIs will conduct a structured internal review assessing: (1) whether testbed quality meets the completeness and coverage criteria defined in **Task 2**; (2) whether the agent configurations tested in Year 2 demonstrate measurable improvement in Groundedness and Fidelity scores relative to a no-agent baseline; and (3) whether practitioner feedback from Year 2 webinars surfaces hardware or usability barriers that require design changes before Year 3 benchmarking. The Advisory Board’s semi-annual meetings serve as an independent checkpoint, with board members reviewing interim findings against the project’s stated RQs and flagging methodological concerns before the next phase begins.

Summative Evaluation. At project close, success will be assessed against four measurable criteria tied directly to **RQ1–RQ4**: (1) the adversarial Verification Agent demonstrably achieves hallucination reduction $> 40\%$ vs. a standard RAG baseline on the shared testbed (**RQ1**); (2) the Auditing Toolkit’s automated Groundedness and Fidelity scores achieve provenance linking $\geq 85\%$ reviewer acceptance (**RQ2**); (3) these scores correlate significantly ($r \geq 0.75$) with expert panel

ratings, and the Extraction Agent achieves metadata extraction precision $\geq 90\%$, which professional archivists rate as audit-worthy (**RQ3**); and (4) benchmarking results yield actionable, quantified data on performance and privacy tradeoffs between locally hosted and commercial models under matched conditions (**RQ4**).

3.5 Project Management

3.5.1 Project Team

Dr. Yinlin Chen will serve as Principal Investigator and Project Director, providing overall leadership and research oversight. An Associate Professor in University Libraries and Associate Director of the Center for Digital Research & Scholarship, he holds a Ph.D. in Computer Science & Applications. His research focuses on generative AI [38], machine learning for scientific and biomedical applications [39, 40], and the development of intelligent frameworks for discovering and managing complex, multimodal digital collections [41–43]. With a strong publication record in leading AI and digital library venues, Dr. Chen manages the project’s strategic execution, supervises the Graduate Research Assistants (GRAs) and a student wage worker, and coordinates the research activities between the Virginia Tech and Northwestern University teams.

Dr. Chang-Tien Lu will serve as Co-PI and lead the AI research components of the project. An Associate Director of the Sanghani Center and Professor of Computer Science at Virginia Tech, he is an ACM Distinguished Scientist and IEEE Fellow with over 200 peer-reviewed publications and extensive interdisciplinary research leadership experience [44–47]. He will define the AI vision and guide technical direction, overseeing experimental design, data strategy, model development, system integration, and comparative evaluation against commercial cloud models, and supervision of GRAs. He will establish rigorous methodological standards to ensure transparent workflows, reliable evaluation, and a reproducible, sustainable toolkit for library and archival applications.

Dr. Kelsey Rydland will serve as Co-PI. Dr. Rydland holds a Ph.D. in Public Administration and is the Operations Director for Academic Innovation at Northwestern University Libraries. He will coordinate the NUL Digital Products and Data Curation units to lead his institution’s participation in the project and oversee local evaluation activities. He will review research findings, coordinate assigned project tasks, and provide institutional input based on evaluation results. Through testing in a second research library environment, he will help assess the applicability of the AI approach and contribute to refining the research framework [48].

Benn Joseph and **Kelsey A. O’Connell** (Project Staff, Northwestern) provide the archival and digital preservation expertise essential for the study’s independent review cycles. Mr. Joseph, Head of Collections Services, focuses on aligning system logic with professional archival metadata standards. Ms. O’Connell, a Digital Archivist who has appraised over 18.2 terabytes of born-digital collections, applies her expertise in metadata and appraisal to validate the auditing toolkit’s fidelity and groundedness metrics [49].

3.5.2 Advisory Board

A multi-disciplinary Advisory Board provides external governance and expert validation for the research framework. The board meets semi-annually to evaluate interim findings and ensure the study delivers a replicable, evidence-based asset. **Dr. Sharon M. Leon** [50] validates community adoption pathways for the open-source SLI toolkit. **Dr. Leo Lo** [51] ensures methodological alignment with the 2026 UVA Archival AI Protocol [9] and national standards for AI governance. **Dr. Mario H. Ramirez** [52, 53] evaluates the integration of automated workflows within professional practice standards to assess the effectiveness of human-in-the-loop verification. **Natalie Buda Smith** [54] ensures research findings are generalizable for national-scale discoverability. **Dr. Tyler Walters** [55] evaluates the long-term sustainability of research outcomes within national digital stewardship infrastructures. **Xuemao Wang** [56] provides guidance on institutional governance and strategic alignment within research library environments. Advisory Board member resumes are in Supportingdoc2, and Letters of Support and Participation are in Supportingdoc3.

4 Project Results

This project will produce the SLI Toolkit, a research-validated reference architecture and evaluation suite whose design, performance characteristics, and portability conditions will be empirically established through the three-year study. By integrating modular AI workflows with standardized auditing measures, the project provides an evidence-based pathway for responsible agentic AI adoption within libraries and archives.

Table 1 summarizes the project’s four primary deliverables and their anticipated impact for libraries and archives.

Table 1: Project Deliverables, Impact, and Sustainability

Project Deliverables	Description & Impact	Access & Sustainability
SLI Architectural Blueprint	A documented reference architecture that separates curatorial policy from model choice, allowing institutions to adopt improved AI models over time without redesigning their metadata workflows.	Tagged releases with versioned documentation and reusable architecture/workflow templates; archived preservation package for long-term reuse and citation.
Evaluated Agent Skills & Prompting Typologies	An open library of agentic skills (prompt templates, workflow sequences, and grounding configurations) designed to process complex archival materials and reduce factual errors in metadata creation.	Rights-cleared configuration sets with evaluation notes and example workflows; archived snapshots to support durable access and reuse.
Archival AI Auditing Toolkit	A standardized evaluation toolkit combining automated scoring and an annotated ground-truth corpus to quantify Groundedness and Fidelity in AI-generated archival metadata prior to deployment.	Open scripts and benchmark datasets packaged for independent replication, comparison to future models, and field-wide extension; archived benchmark releases.
Practitioner’s Implementation and Governance Manual	A practical “if-then” adoption playbook addressing hardware requirements, governance controls, and human-in-the-loop oversight for mid-sized and resource-constrained libraries.	Released in preservation-friendly formats (PDF/A + Markdown) with changelog/versioning to support accessibility and iterative community updates; archived editions for long-term preservation.

By structuring all outputs for open release and long-term preservation, the SLI Toolkit is designed as a durable and reusable resource for the national library community. The integration of reference architectures, evaluated agentic workflows, standardized auditing measures, and governance guidance establishes a sustainable foundation for transparent, accountable, and institutionally controlled AI adoption in archival practice.

Dissemination Plan: Workshops, tutorials, and presentations will be delivered at major library and archive conferences (ALA, SAA, ACRL, DLF, CNI) and at conferences in Artificial Intelligence (AAAI, NeurIPS, ICML, CVPR, ACL, IJCAI), Information Retrieval (SIGIR, KDD), and Digital Libraries (JCDL, ICADL, TPD). All project outputs will be released under an MIT License on GitHub and preserved in OSF and VTechWorks.

Budget Allocation: The project budget includes \$32,700 for 5% effort from the PI and Co-PI to design and supervise the agentic AI research and \$7,378 in faculty fringe benefits, while a \$82,737 subaward to Northwestern University supports 5% effort from the co-PI and project staff for independent portability validation. Student Support of \$277,235 funds the GRAs and a student wage worker for research activities. Travel of \$25,250 is dedicated to research dissemination, presentations, and workshops, with \$2,000 for Generative AI benchmarking credits, \$7,500 for advisory board honorariums, and \$178,306 in indirect costs.

Schedule of Completion

Project Period 1: 09/01/2026 - 08/31/2027[illegible]

Schedule of Completion

Project Period 2: 09/01/2027 - 08/31/2028[illegible]

Schedule of Completion

Project Period 3: 09/01/2028 - 08/31/2029[illegible]