



LAURA BUSH 21st CENTURY LIBRARIAN PROGRAM

Sample Application RE-260213-OLS-26
Project Category: Early Career Research

Purdue University Libraries and School of Information Studies

Amount awarded by IMLS: \$421,376
Amount of cost share: \$0

In this Early Career Research project, Dr. Spencer Dean Stewart at Purdue University School of Information Studies will investigate restoring lost structure to digitized historical records. While massive digital archives exist, they often get flattened during digitization, lacking references to speakers in transcripts or recipients in letters. This project will use Generative AI to create a structural recovery framework that automatically identifies these relationships while preserving archival accuracy and transparency. In partnership with Purdue University Archives and Special Collections and a team of external disciplinary consultants at major U.S. institutions, the project will create a generalizable and reproducible AI workflow, deploy it on three case studies, and produce online modules to make the tools more accessible to others.

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.

Archival Intelligence: Reproducible AI Frameworks for Structural Recovery

I. INTRODUCTION

The Purdue Libraries and School of Information Studies requests \$421,375 in IMLS LB21 funding for a two-year Early Career Research project led by Spencer Dean Stewart to support *Archival Intelligence: Reproducible AI Frameworks for Structural Recovery*. This project addresses a national challenge in libraries and archives: restoring lost structure to digitized historical records. While librarians, archivists, and scholars recognize that highly structured, marked-up datasets (like TEI-XML) are useful for accessibility and research, the manual labor required to produce them limits their application to small, boutique corpora. Our framework for **Structural Recovery** uses Generative AI to scale up this process by reclaiming the semantic and functional relationships inherent in archival documents that are frequently flattened during the OCR process. These include creating marked-up corpora that restores essential features from semi-structured data, the formal layout of correspondence (such as dates, recipients, or addresses), and speaker identification in congressional records. Through case studies involving U.S. historical documents, we will produce validated pipelines, benchmark datasets, and professional training modules that will test Generative AI's ability to assist in libraries and archival work and develop shared standards for using AI in a way that preserves archival values such as transparency and provenance (Lo 2026). In partnership with Purdue University Archives and Special Collections and external disciplinary consultants, this project addresses the following research questions related to design, validation, and generalizability: (RQ1) What design principles are needed to enable AI-assisted workflows to reconstruct lost textual structures while still preserving key archival values such as provenance, transparency, and accuracy? (RQ2) What quantitative and qualitative validation methods are necessary to benchmark the accuracy and reproducibility of AI-assisted archival transformations? (RQ3) How transferable are structural recovery workflows across distinct data structures and institutional contexts? By focusing on specific case studies, we aim to create AI systems that value the core missions of libraries and archival practices as grounded in the materiality of everyday documents.

II. PROJECT JUSTIFICATION

Problem Space: Digitization without Structural Accessibility

Over the past two decades, libraries and archives have made significant investments in digitizing historical collections. Millions of pages of legislative records, administrative correspondence, scientific reports, and printed inventories are now available online. As of February 2026, the U.S. National Archives alone have digitized more than 455 million pages of records. While this represents a massive step forward in accessibility and preservation, digitization has not resolved many of the deeper structural barriers that limit meaningful access. For example, digitized texts are frequently stored as flattened corpora that obscure internal organization, such as individual speakers within debates or discrete entries within tabular records. Archival scholarship has long recognized tensions between scale and descriptive depth. Greene and Meissner's (2005) influential "More Product, Less Process" (MPLP) framework argued for reducing intensive item-level processing in order to address backlogs and increase access at scale. Subsequent discussions extended these concerns into digitization contexts, noting

that while large-scale digitization improves baseline access, it often does so at the expense of granular descriptive structure (e.g. Sutton 2012).

Past efforts to bridge this gap have faced significant hurdles. Traditional Natural Language Processing (NLP) often struggles due to both the "noisy" text produced by historical OCR and the reality that NLP systems struggle with historical materials due to the fact that they are trained on contemporary data (Gladstone et al., 2025). Other attempts at restoring structure such as the Text Encoding Initiative (TEI) are robust, but they remain slow and laborious (Ide & Veronis, 1995). These manual or semi-manual workflows have largely been restricted to "canonical" or well-funded digital editions, leaving the vast majority of archival records in a state of digital limbo (Flanders 2016). Consequently, researchers can view scanned pages but cannot easily query, aggregate, or analyze the granular units of information trapped within them. This divide between mere digitization and true structured accessibility is a pressing challenge for institutions seeking to unlock the full public and scholarly value of their collections.

Recent advances in Generative and Agentic AI systems present promising solutions to some of these problems. Recent research on large language models (LLMs) have proven useful for enhancing traditional OCR and Handwritten Text Recognition (HTR) pipelines (Crosilla et al., 2025; Poznanski et al., 2025). However, transcription accuracy alone does not resolve the deeper structural problem. Many historical sources such as legislative debates, inventories, and scientific reports contain layered internal schema that are not captured by flat text extraction. Recovering speaker-level units, tabular rows, or relational metadata requires reconstructing structural relationships, not merely recognizing characters. While there is a growing body of literature demonstrating the potential of Generative AI systems for information extraction (Ma et al., 2024; Stuhler et al., 2025; Polak & Morgan 2024; Alizadeh 2025) and Document Layout Analysis (Binmakhashen & Mahmoud 2019; Sun et al. 2025), including my own work (Stewart & Sinha 2025; Stewart & Butler, forthcoming), many of these implementations have a main limitation: they remain collection-specific prototypes with limited documentation of validation protocols, reproducibility standards, and infrastructural transferability. As such, they do not necessarily preserve archival and library values such as provenance, transparency, and traceability.

Together, these strands of scholarship reveal an institutional gap: while libraries have developed strong standards for description and made significant progress in digitization and transcription, there remains no widely documented framework for reconstructing granular archival structure in ways that are reproducible, validated, and scalable across institutional contexts.

Intervention: A Reproducible Framework for Structural Recovery

This project introduces **structural recovery** as a methodological response to the limitations of digitized archives. Structural recovery refers to the reconstruction of granular schema that is present in the original materials but obscured in digitized form. Rather than treating digitized pages as static images or undifferentiated text blocks, structural recovery transforms them into machine-readable, relationally structured datasets while maintaining traceable links to original archival identifiers. To demonstrate the portability of this framework, we apply it to three distinct archival challenges. Holding the architecture constant across these cases ensures each functions as a controlled comparative environment rather than a standalone technical demonstration.

Case Study 1: Flattened Discourse (U.S. Congressional Record, 1873-Present) The Congressional Record exists largely as flattened transcripts where speaker contributions are embedded within continuous text. This structure prevents granular analysis of individual speakers.

- Recovery Task: Delimit discrete speaker turns and normalize fields (names, party, state, bill identifiers).
- Impact: Transforms “stream-of-consciousness” transcripts into longitudinal, speaker-level datasets suitable for analysis.

Case Study 2: Hybrid Tabular-Narrative (USDA Plant Inventory, 1898-Present) These volumes document global agricultural development through a mix of prose, tables, and marginalia. Standard digitization fails to expose the entry-level schema.

- Recovery Task: Reconstruct entry boundaries and extract specific fields (accession numbers, botanical names, origin locations, and dates) using historical gazetteers for normalization.
- Impact: Turns a century of agricultural history into a queryable database of global biodiversity and seed movement.

Case Study 3: Fragmented Discovery (Purdue University Barron Hilton Archives for Flight and Space) The collections within the Barron Hilton Archives for Flight and Space Exploration at Purdue University Archives and Special Collections, including the Neil A. Armstrong papers and related aeronautics and astronautics materials are described through hierarchical finding aids but relational connections among missions, laboratories, and individuals are dispersed and inconsistently catalogued across digitized collections.

- Recovery Task: Extend metadata extraction from disparate archival materials such as letters, speeches, and reports to include relational units such as people, programs, dates, missions, places, etc.
- Impact: By extracting and normalizing these entities, we construct a relational index that links individual items (e.g., a letter in one box to a mission report in another collection), facilitating cross-collection discovery that maintains original archival order.

Across all three cases, this project builds on existing scholarship by: focusing on recovering archival structure rather than merely improving search; integrating human-in-the-loop validation protocols to safeguard key values in archives and libraries studies such as accuracy and transparency (Lo 2026); and producing documented, transferable and scalable workflow standards rather than isolated technical demonstrations.

Beneficiaries:

This project delivers high-impact benefits across three primary stakeholder groups:

- **Libraries and Archives:** The project empowers institutions, particularly small-to-mid-sized academic and regional archives, to scale their digital accessibility. By providing established workflows, it bridges the gap for organizations that face rising expectations for optimized digital access but lack dedicated in-house AI research teams.
- **Researchers:** Scholarly communities will gain access to structured, analysis-ready datasets. These resources enable granular, data-driven exploration of diverse topics, transforming raw digitized images into searchable, high-utility research assets.

- **The Broader Community:** Public users will benefit from enhanced discoverability of historical materials vital to public policy and governance. In alignment with the United States' 250th anniversary, the project unlocks foundational records of American governance, scientific progress, and agricultural transformation, fostering civic engagement during a landmark period of national reflection.

By establishing reproducible workflows and rigorous validation standards, this initiative lowers the barrier to responsible AI adoption across the national cultural heritage sector.

Long-term Research Agenda:

Dr. Stewart's long-term research program examines how computational methods can transform large-scale historical corpora into structured, analyzable, and publicly accessible knowledge infrastructures. The emphasis on structured knowledge infrastructures aims to address the longstanding problem in computational research where keyword searchability often comes at the expense of archival context and structural integrity (Putnam 2016). His prior work has focused on testing the capabilities of Generative AI systems for tasks such as speech segmentation, information extraction, geospatial reconstruction, and relational metadata modeling across diverse multilingual datasets (Stewart & Sinha 2025; Stewart & Butler forthcoming; Gladstone et al. 2025; Fang et al., 2025). These projects have demonstrated the analytical value of converting flattened textual archives into structured datasets. This Early Career project marks a strategic evolution from disciplinary-specific case studies toward the creation of generalizable methodological frameworks. The focus moves beyond building bespoke pipelines for individual research questions to formalize the principles of structural recovery for broad library and archival implementation.

Moving forward, Dr. Stewart has two main research agendas: develop generalizable frameworks for structural recovery and utilize these tools to advance historical inquiries. The core of this research grant aims to address this first item by resolving methodological challenges. It aims to help move libraries and archival practices beyond the searchability trap towards a structural model of preservation that reconstructs the relational logic of the original record (e.g. how a single letter relates to other collections, or the relationship between a congressional speaker and a bill). Formalizing structural recovery pipelines will allow libraries and archives to unlock these hidden forms of data in their collections. Moving forward, his future research will utilize these newly unlocked materials to answer specific historical questions. These will include a large-scale study of the origins of American agriculture through an analysis of datasets like the USDA Plant Inventory records included in this study, allowing us to map the global movement of seeds and the rise of American agriculture. It will also include a study of the architecture of space exploration through the unlocking of fragmented archives at Purdue to reconstruct the relational networks of engineers, laboratories, and policy-makers that fueled the Space Age. These research agendas ultimately seek to ensure that as the record of human history becomes increasingly algorithmic, it remains transparent, contextual, and authoritative. By bridging the gap between high-level computational methods and the traditional rigors of history and archival science, Dr. Stewart aims to contribute to a future where we no longer just "search" history through keywords, but instead take advantage of the affordances of digital technologies to navigate its complex, relational architecture.

III. PROJECT WORK PLAN

This project is grounded in traditional theories in archival and libraries practices, including the growing field of computational archival science (e.g. Hedges et al., 2022). We aim to bridge the gap between “flat” digitized text and “deep” archival intelligence. The project addresses the following research questions:

- **RQ1 (Design):** What architectural principles enable AI-assisted workflows to reconstruct lost archival schema while preserving key archival values such as provenance, transparency, and accuracy?
- **RQ2 (Validation):** What quantitative and qualitative validation methods are necessary to benchmark the accuracy, reproducibility, and reliability of AI-assisted archival transformations?
- **RQ3 (Generalizability):** How transferable are structural recovery workflows across distinct archival data structures and institutional contexts?

The primary goal for Year 1 is to establish a solid foundation for the initiative by developing and prototyping a unified structural recovery framework tailored for archival environments. During this phase, the project will focus on material development, internal pilot testing, and early stakeholder engagement to ensure that the recovery workflows are accessible, effective, and fully aligned with real-world library standards (RQ1). This groundwork will prepare the initiative for robust implementation and comparative analysis beginning during the latter part of Year 1 and into Year 2 (RQ2 and RQ3).

Activity 1: Team Formation and Advisory Board Recruitment: The first phase of Year 1 will begin with a project kickoff designed to establish an agile framework that allows for iterative progress and continuous improvement. The PI, an Assistant Professor of Digital Humanities with expertise in computational analysis and historical informatics, will provide overall project leadership. The PI will be assisted by a Graduate Assistant with proficiency in Natural Language Processing and Generative AI. Together, they will manage the modular four-part workflow (segmentation, extraction, provenance, and validation). Specific roles will be assigned, and clear communication channels will be established with shared file folders. Simultaneously, the team will formalize the advisory board, involving stakeholders from both Purdue University and other academic libraries to ensure the project reflects actual user needs. Success Metrics will be measured by the timely formation of the project team and advisory board and the creation of a comprehensive, milestone-driven project plan.

By the time the proposal was submitted, the following librarians and experts have committed to participate on the Advisory Board. The support letters are included in the supporting documents accompanying this proposal.

Name	Role & Expertise
Gang Shao, Ph.D.	Associate Professor, Purdue University. Expert in AI/Data science and GeoAI.
Jo Otremba, MLS	Barron Hilton Archivist for Flight and Space Exploration and

	Clinical Assistant Professor, Purdue University. Expert on archival theory and practice, aeronautics and astronautics archival collections.
Le Yang, Ph.D.	Associate Vice Provost and Senior Librarian, University of Oregon. Senior library professional & metadata.
Jeffrey Turner, Ph.D.	Assistant Professor, University of Alabama. Historical consultant & expert of congressional records.
Clovis Gladstone, Ph.D.	Associate Director of the ARTFL Project, University of Chicago

Activity 2: Framework Calibration and Pilot Testing: This activity will focus on creating a holistic framework for structural recovery across our different case studies (RQ1). It will outline our approach to creating “ground truth” datasets using our different records, with attention to the different layers of data and metadata needed to ensure persistent links are maintained between original source materials and the AI output. During this phase, we will identify the specific open-source (e.g., Llama, Mistral) and commercial (e.g., GPT-4, Claude) Large Language Models (LLMs) to be used and begin drafting prompt strategies for structural segmentation and entity extraction. We will further develop a robust evaluation pipeline to check for LLM hallucination and establish baseline precision/recall for rule-based heuristics.

Success Metrics:

- Completion of a set of guidelines and sample outputs for feedback
- Established baseline accuracy metrics for pilot datasets
- Finalized evaluation pipeline for hallucination detection
- Selection of LLM candidates and initial library of tested prompt templates.

Activity 3: Advisory Board Engagement and Initial Feedback: Once a framework has been developed, the board will convene its first meeting to review the project work plan, discuss early findings, and offer strategic recommendations. Their input will be carefully documented and integrated into the framework and future plans for the project. These sessions are key to ensuring our framework upholds best practices in the libraries and archives spaces.

Success Metrics:

- 100% participation from committed board members
- Documentation of strategic pivots based on advisor input
- Formal validation of the framework's alignment with library standards.

Activity 4: Case Study 1 Implementation (Flattened Discourse): Full deployment of the workflow for U.S. Congressional Record (1873-present) to reconstruct internal debate structure. This focuses on transforming flattened transcripts into granular, speaker-level units.

Success Metrics:

- Generation of a functional processing pipeline including refined prompt strategies for speaker identification
- Creation of a ground truth benchmark for longitudinal legislative debate

- Comparative evaluation of performance across selected open-source and commercial LLMs
- Documentation of a “Best Practices” series for segmenting flattened archival text
- A finalized longitudinal speaker-level dataset that meets our target accuracy threshold

Activity 5: Case Study 2 Implementation (Tabular-Narrative): Execution of the recovery workflow on the USDA Plant Inventory volumes, addressing the challenge of shifting layouts that combine prose and tabular fields.

Success Metrics:

- A refined pipeline for entry-boundary detection and entity normalization in hybrid layouts
- Creation of ground truth dataset for botanical and geographic entities
- Performance report comparing LLM efficiency in handling structural ambiguity in tabular-narrative text
- Validated best practices for maintaining data provenance in complex page layouts
- A finalized structured dataset that meets our target accuracy threshold

Activity 6: Case Study 3 Implementation (Fragmented Discovery): Application of structural recovery across Flight and Space related archival materials at Purdue. This activity identifies relational units (individuals, missions, programs) across fragmented finding aids and digitized items.

Success Metrics:

- An extraction pipeline optimized for machine-aggregated relational metadata
- Benchmarks for cross-collection entity resolution (linking individuals to specific metadata links)
- Evaluation of LLM performance in metadata extraction and normalization
- Best practices for structural recovery within provenance-based discovery systems
- A finalized structured dataset capable of performing cross-collection analyses

Activity 7: Systematic Validation and Reproducibility Audits: Once we have completed all of our case studies, we will then conduct a formal validation across all cases (RQ2). This includes precision/recall measurement, entity-level validation, and held-out reproducibility audits to analyze architectural strengths and limitations. Tasks include: publication of a comprehensive Validation Report summarizing performance across all case-study pipelines; the completion of cross-case comparative benchmarks; a finalized "Model Selection Guide" based on the performance audits from previous activities; and guidelines for prompt engineering.

Activity 8: Finalization of Materials and Framework Refinement: Building on the insights gathered during case study implementation, the advisory board will again meet to review the findings and suggest further framework refinement. During this activity, the structural recovery pipelines and documentation will undergo iterative revisions to enhance clarity, functionality, and accessibility. With regard to refining workflows, we will incorporate additional edge cases identified during the different case studies to better illustrate AI applications within varied

archival workflows. We will also finalize instructional materials with universal design principles and multimodal content to support diverse institutional needs.

Activity 9: Impact Evaluation, Toolkit Release, and Sustainability Planning: The final phase of the project focuses on a comprehensive evaluation of the framework's impact and the public release of the "Structural Recovery Toolkit". The release of the 'Structural Recovery Toolkit' serves as the primary deliverable for RQ3, providing the library community with a model-agnostic, scalable framework that can be adapted to diverse institutional budgets and technical capacities. Throughout the period, the team will maintain a focus on performance metrics and usability data to assess the curriculum's overall effectiveness in meeting the needs of diverse librarian audiences. To ensure long-term sustainability, we will implement a structured management plan using GitHub as a central hub for managing and updating the recovery materials, allowing the project to leverage version control for transparent updates and open-source contributions from the broader community. Success in this activity will be measured by the release of the open-access toolkit containing refined pipelines and prompt libraries, the successful delivery of professional workshops, and the publication of a "Library Best Practices for Structural Recovery" guide.

IRB Statement

This project does not involve human subjects research. All materials consist of publicly accessible or archival historical records. No personal data will be collected. Institutional Review Board approval is therefore not required.

IV. PROJECT RESULTS

Intended Results and Deliverables

The project will generate a suite of resources designed to move AI from an experimental curiosity to an operational archival tool. The primary outputs include:

- **The Structural Recovery Framework & Toolkit:** A modular, model-agnostic "field manual" for AI-assisted recovery. This includes optimized pipelines for segmentation and extraction, accompanied by a library of prompt templates compatible with both local/open-source and commercial LLMs.
- **Three Validated Benchmark Datasets:** Three distinct "proof-of-concept" datasets—longitudinal legislative records (U.S. Congressional Record), hybrid scientific tables (USDA Plant Inventories), and fragmented discovery metadata (Purdue Flight and Space Collections).
- **Validation and Benchmarking Standards:** Publicly documented protocols for auditing AI-generated metadata, including error typologies and "hallucination" detection strategies tailored for archival environments.
- **Professional Development Modules:** A curriculum of recorded workshops and instructional materials designed to foster AI literacy and institutional capacity across the library and archives profession.

Addressing the Identified Need: This project directly confronts the "structural gap," where millions of historical records are digitized but functionally dormant to researchers and the public

because they remain in flattened or semi-structured formats. By providing a validated, provenance-sensitive pathway to reconstruct granular schema, we move beyond mere digitization toward structured accessibility. These results enable researchers to perform more precise queries and allow institutions to enhance the value of their collections without the prohibitive costs of manual re-cataloging.

Adaptability, Generalizability, and Usability: To ensure the project serves the broadest possible community, our framework is built on a philosophy of institutional flexibility. Rather than releasing static scripts, we will provide modular components that help archivists estimate the time and computational resources required for implementation based on their specific technical maturity. By stress-testing the framework against three wildly different data structures—narrative, tabular, and fragmented—we ensure that the resulting guidelines are transferable to diverse archival genres. All code, documentation, and validation protocols will be released through a public repository under an open license, ensuring that a regional archive with a modest budget can achieve results comparable to a large research university.

Dissemination Strategy: Our dissemination strategy focuses on meeting library and information science professionals where they already work. We will present our findings at major national venues, including the Association of College & Research Libraries (ACRL), the Digital Library Federation (DLF) and Society of American Archivists (SAA), while publishing peer-reviewed research in key venues for library and archival science. Beyond formal publication, we will host virtual workshops through Purdue’s Libraries and School of Information Studies to provide hands-on training. To ensure long-term discovery, all digital assets will be assigned permanent DOIs and hosted on both Purdue’s online data repository (PURR) and GitHub, with mirrors in stable institutional repositories.

Sustainability Beyond the Grant Period: The benefits of this project will be sustained through a dual focus on institutional integration and community stewardship. As Director of the Purdue Digital Humanities Studio, the PI will incorporate these workflows into the Studio’s permanent service infrastructure, ensuring they support future faculty and student research. Furthermore, the training modules will be woven into the PI’s graduate-level curriculum in information studies, training the next generation of AI-literate librarians. By emphasizing a model-agnostic, open-source approach, the framework avoids dependency on specific vendors, allowing the broader community to iterate and update the toolkit as AI technologies continue to evolve.

Schedule of Completion

[illegible]