



NATIONAL LEADERSHIP GRANTS FOR LIBRARIES

Sample Application LG-260016-OLS-26
Project Category: Planning

Indiana University

Amount awarded by IMLS: \$154,419
Amount of cost share: \$0

The Ostrom Workshop at Indiana University and the Indiana University Luddy School of Informatics, Computing, and Engineering, in partnership with Indiana University Libraries, will plan to design and evaluate a proof-of-concept for artificial intelligence-enabled access and stewardship of heterogeneous social science databases in libraries and archives. The direct beneficiaries of this project are research libraries seeking pathways to integrate artificial intelligence into structured data stewardship, specialized research centers stewarding complex databases but lacking scalable interface solutions, and data librarians and research data services units navigating artificial intelligence adoption. By establishing a governance framework and technical model, the project seeks to equip these institutions to lower barriers to complex research data, ultimately serving the educators, interdisciplinary scholars, graduate students, and civic learners who engage with questions of community governance, institutional design, and shared resource management but currently lack meaningful access to the empirical record these databases contain.

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.

AI-Enabled Access and Stewardship for Heterogeneous Social Science Databases in Libraries and Archives

Introduction

The Ostrom Workshop at Indiana University (IU) and the IU Luddy School of Informatics, Computing, and Engineering, in partnership with IU Libraries, request \$181,424 from the Institute of Museum and Library Services to support a two-year Planning project to design and evaluate a proof-of-concept (POC) for AI-enabled access and stewardship of heterogeneous social science databases in libraries and archives. The project aligns with Objective 1.3 as it will enhance information infrastructure and promote access through emerging technologies.

Social science databases built through decades of research remain functionally inaccessible due to legacy technical architectures, limited user interfaces, and the specialized expertise required to query them, leaving critical components of our information infrastructure underutilized. **This project will assess the feasibility of using AI to reduce these barriers**, producing: (1) a governance and stewardship framework for AI mediation of research databases; (2) a technical proof-of-concept demonstrating natural language access across heterogeneous databases; (3) a needs assessment with design specifications and recommendations adaptable by libraries and research data stewards nationally.

Project Justification

Social science databases constitute a significant part of the nation's scholarly heritage, capturing empirical evidence about how societies and communities create, organize, and use their resources, and respond to social and environmental change. These databases form a foundational corpus for social science and civic inquiry in the US (Scheuch, 2003). Often initiated through federal seed funding, they are housed across research centers and local environments rather than systematically integrated into university and library infrastructures. Such diversity of institutional homes results in uneven governance and funding and sustainability models, creating significant barriers to discovery, interoperability, and long-term access. As the knowledge embedded in these collections is relevant to educators, policy practitioners, and civic learners, their inaccessibility is not only a scholarly loss but a public one.

The complex and distributed governance and technical architectures of these databases translate directly into barriers to meaningful use. Research on data reuse consistently demonstrates that complexity, insufficient documentation, and the need for specialized query expertise significantly reduce secondary analysis and interdisciplinary uptake (Curty et al., 2017; Faniel & Yakel, 2017). Even when databases exist in stable digital environments and have basic discovery mechanisms, they often require extensive interpretation of codebooks, schema relationships, and variable definitions before a single research question can be executed. Effective use frequently depends on insider knowledge or advanced technical skills, particularly the ability to construct and validate structured queries across multiple relational tables. As a result, data that is nominally discoverable remains functionally inaccessible to researchers, educators, and students.

The examples examined and summarized below illustrate this pattern. Each database has diverse established management and maintenance models, yet each poses significant and recurring challenges in access, schema interpretation, cross-database comparison, or user-facing interface design. Together, they demonstrate that technical access does not resolve the deeper issues of interpretive and functional access.

Database & URL	Management	Maintenance model	Access & user challenges	Library involvement
Varieties of Democracy v-dem.net	Multiple academic institutions across the US and EU	Experts code indicators	• 600+ indicators; highly complex codebook • Many variables are latent constructs, not raw data • Coder-level data in separate archive	Active (U of Notre Dame Libraries archives dataset and provides access)
Comparative Constitutions Project comparativeconstitutionsproject.org	Non-profit, led by UT Austin	Small RA/postdoc team	• 650+ characteristics coded at high abstraction • Codebook literacy essential • Schema alignment needed to merge with other datasets	Basic (U of Illinois digitization infrastructure, Harvard Dataverse for downloads)
Database of Religious History religiondatabase.org	Academic (U of British Columbia)	Expert contributions, editors curate tags and recruit experts	• Extensive ontology; entry granularity varies by expert • Nested data structure • Statistical use requires data flattening expertise	Active (UBC institutional repository stores and provides access)
Land Matrix Initiative landmatrix.org	Multi-stakeholder consortium	NGO and government field research + crowdsourcing	• Data quality and completeness vary significantly by country • Complex schema (investors, land area, use type, community impacts, conflicts)	No significant library involvement
Atlas of Economic Complexity atlas.hks.harvard.edu	Academic (Harvard U)	Team of researchers	• Complexity indices (ECI, PCI, COI) are derived metrics requiring methodological understanding • Atlas tool vs. downloadable Dataverse data require separate navigation • Confusion with OEC (MIT spin-off), a distinct commercial product	Basic (Harvard Dataverse and Harvard University Libraries hosts some datasets)

Across these examples, libraries rarely play a strategic or integrated role, and mechanisms for interpretive access and cross-collection usability are absent. Even when databases are hosted on stable servers with documentation, access typically assumes familiarity with internal schema logic, variable naming conventions, and historical data collection practices, making exploratory or cross-cutting inquiries difficult. This challenge extends to formal repositories: a search of ICPSR, the world's largest social science data repository, returns over 3,600 datasets categorized as databases, with files distributed across heterogeneous formats and requiring users to construct their own relational environments before analysis can begin.

Current hosting arrangements of social science databases, therefore, present two different problems: they either prioritize maintenance and data integrity without mechanisms for managing metadata and rights and pursuing longevity or they focus on preservation and require users to download files and create their own environments. Either way, there is an observable lack of scalable mechanisms for lowering the cognitive and technical barriers to using such databases (Borgman et al., 2015; Faniel & Yakel, 2017; Gregory et al., 2020). The planning framework developed through this project is thus applicable not only to databases outside formal library or repository systems, but to the broader challenge of making complex, structured research data accessible.

Recent advances in AI, particularly in natural language interfaces to structured databases and large language model-driven semantic parsing, offer a potential pathway for mediating access across heterogeneous schemas (Ma et al., 2024; Yu et al., 2019; Zhong et al., 2017). However, these systems are benchmarked on standardized, well-documented database environments and perform poorly when applied to historically layered research databases with complex schemas and evolving documentation (Li et al., 2023). Moreover, research on hallucination and reliability in AI underscores the need for validation mechanisms specifically adapted to institutional research contexts (Huang et al., 2025; Ji et al., 2023). No current tool is designed for seamless natural language access to this type of data.

Academic libraries possess the complementary expertise needed to fill this gap, including metadata standards, user-centered access design, digital preservation, and cross-collection interoperability. With such expertise, they are well positioned to serve as institutional anchors, yet no tested model currently integrates library governance oversight with AI-enabled access and institutional stewardship of research data into a replicable framework. Such an effort also aligns with federal policy priorities that emphasize both accelerating AI innovation and strengthening institutional capacity for responsible and educationally grounded AI adoption (*Ensuring a National Policy Framework for Artificial Intelligence*, 2025).

The proposed Planning project addresses this gap by (1) defining governance and stewardship constraints for AI-enabled access, (2) designing grounded architecture for natural language mediation across heterogeneous databases, and (3) evaluating feasibility through a limited proof-of-concept. We selected three databases as our test case and will convene a diverse group of stakeholders, including librarians, data stewards, domain researchers, and technical experts with interest in long-term stewardship of these collections. By structuring the work around specific databases and documented use cases, the project will generate insights about whether AI can enable natural language access across multiple socially and historically relevant databases and when and how AI mediation is appropriate in library-hosted research environments. While valuable as a standalone contribution to library and archival practice that will strengthen access and stewardship of social science data, this Planning effort is designed as the first phase of a larger, future multi-institutional initiative that will create a model for using AI in research data environments.

The direct beneficiaries of this project are research libraries seeking pathways to integrate AI into structured data stewardship, specialized research centers stewarding complex databases but lacking scalable interface solutions, and data librarians and research data services units

navigating AI adoption. By establishing a governance framework and technical model, the project equips these institutions to lower barriers to complex research data, ultimately serving the educators, interdisciplinary scholars, graduate students, and civic learners who engage with questions of community governance, institutional design, and shared resource management but currently lack meaningful access to the empirical record these databases contain.

Project Work Plan

This planning project builds on two years of our prior preservation and stewardship work on three historically significant social science databases created by Nobel-Prize-winning economist Elinor Ostrom and collaborators (Frey et al., 2016; Ostrom, 1990; Poteete & Ostrom, 2003).

1. The Common Pool Resource (CPR) database, developed in the 1980s, contains systematically coded analyses of 86 case studies of social-ecological systems, capturing at least 40 person-years of prior fieldwork by economists, political scientists, sociologists, anthropologists, and engineers across fisheries, irrigation systems, forests, and grazing lands.
2. The International Forestry Resources and Institutions (IFRI) database, established in 1992, documents forest governance and sustainability across more than 350 communities and 9,000 forest plots in multiple countries, representing one of the largest longitudinal datasets of small-scale forest governance in existence.
3. The Nepal Irrigation Institutions and Systems (NIIS) database, one of the largest irrigation databases ever established for Nepal, covers 233 irrigation systems across 29 districts, capturing institutional, operational, and infrastructural dimensions of water governance using the Institutional Analysis and Development framework.

Together, these three databases constitute an irreplaceable empirical record of how communities govern shared natural resources, and are relevant to contemporary questions of civic governance and resource management. Yet the databases represent a significant challenge to access and use. Collectively comprising over 35 interlinked tables and hundreds of variables, spanning biological conditions, infrastructure measurements, and abstract governance principles, they capture deeply layered relationships between resource systems, communities, and institutional rules across dozens of countries and multiple decades. Answering even a straightforward research question requires knowing which tables to consult, how they relate to each other, and how variables were defined and recorded across different field sites and time periods. This complexity places meaningful use out of reach for anyone without specialized technical expertise and deep familiarity with the collections.

We have already undertaken foundational preservation activities for these databases, including conversion of legacy formats into an open-format MySQL relational database, initial database audits, schema mapping across versions, and a preliminary legal and ethical review of data governance considerations. The team has also deployed these databases onto a university server environment and implemented a web-accessible search interface enabling keyword-based discovery across the collections. These efforts have stabilized the collections, establishing the foundation for the governance, stakeholder, and technical work proposed here.

The project is organized into four interrelated phases: 1) Partnership Structuring and Governance, 2) Needs Assessment, 3) Proof-of-Concept, and 4) Testing and Dissemination (see Figure 1 below).

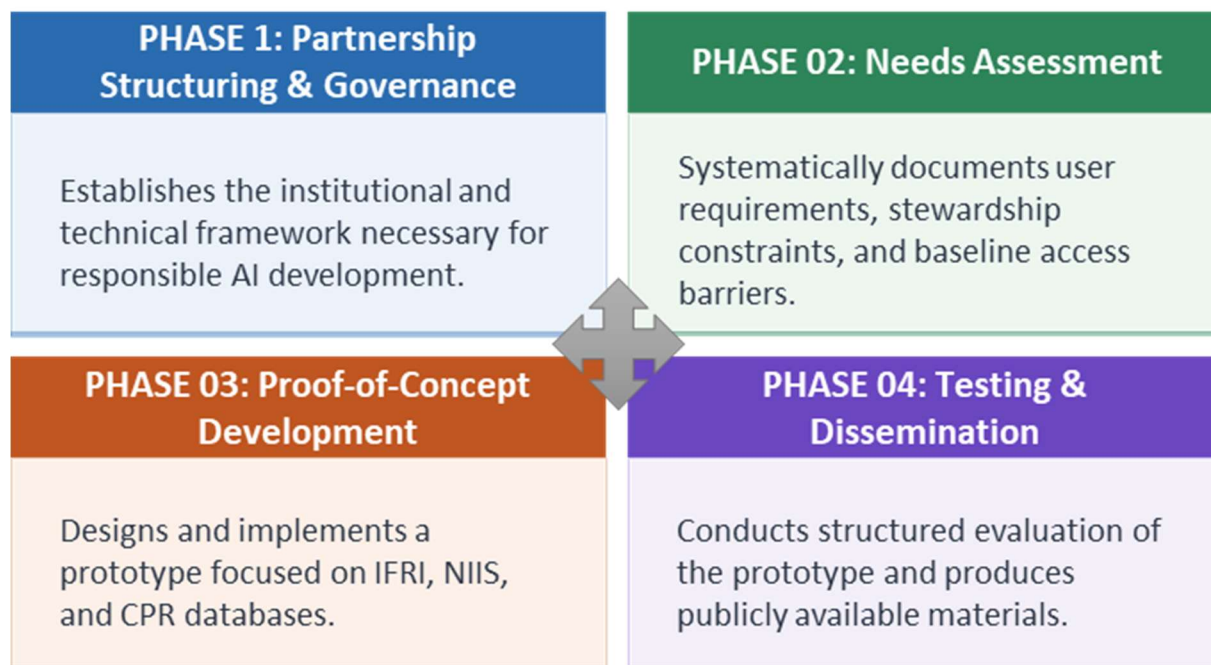


Figure 1. Four phases of the project.

Phase 1: Partnership Structuring and Governance Design

The project will begin with formalizing and expanding the partnership network necessary for responsible AI-enabled access. The team has already established working relationships with IU Libraries Research Data Services, IFRI and NIIS database stewards from the University of Notre Dame and the Indian School of Business and researchers who can serve as users of these collections (see letters of support). We will extend this network through outreach to the SES library at Arizona State University, which hosts an instance of the CPR database and to other known stewards maintaining instances of the NIIS and IFRI databases¹. This outreach will allow us to document the number, location, format, and governance conditions of existing copies across institutions, establishing the full scope of the collections and surfacing schema variations and documentation differences that must be accounted for in subsequent technical work.

Parallel to the outreach, the team will conduct a structured review of current AI approaches to database querying, including natural language-to-SQL methods and model evaluation practices, with attention to transparency, explainability, and error handling in heterogeneous environments. Findings from this review will directly inform the technical constraints embedded in the governance charter.

The primary deliverable of Phase 1 will be a documented collaboration framework and governance charter. Rather than imposing unified control over independently maintained copies, the charter will establish lightweight coordination mechanisms: agreements among stewards to document and share schema differences, notification protocols when any institution modifies or extends their copy, and shared attribution and citation standards. The charter will guide

¹ <https://ulrichfrey.eu/en/niis/>

subsequent technical design and ensure that AI integration strengthens, rather than displaces, distributed library-based stewardship.

Phase 2: Needs Assessment

Phase 2 will translate the outreach and background work of Phase 1 into clearly defined user requirements, stewardship constraints, and technical feasibility criteria. The goal is to document what responsible AI-enabled access must accomplish in order to serve specialized data stewards, broader library infrastructures, and research communities. This phase combines an in-person core stakeholder convening, a broader virtual forum, targeted interviews, and systematic analysis of findings.

The core stakeholder event will be held in person at the Ostrom Workshop and will convene domain research users, data stewards and library professionals, and technical experts. Domain participants will include researchers who have previously worked extensively with the IFRI, NIIS, or CPR databases, spanning institutional economics, anthropology, and political science, as well as doctoral students and policy-oriented researchers. The convening is designed to incorporate critical scrutiny of AI assumptions alongside enthusiasm, ensuring that governance constraints and failure modes receive equal attention to technical possibilities. Library representation will include a research data librarian with expertise in metadata and repository services and a professional with experience in cross-collection interoperability. Technical participants will include the project AI developer, a database expert, and a usability specialist. A data governance expert will participate to clarify governance and modification constraints.

The in-person event will begin with a structured demonstration of current database workflows to make visible the friction points associated with heterogeneous schemas and SQL-dependent access. Participants will work through exercises designed to surface specific usability barriers. The group will then map the governance landscape, identifying data stewardship boundaries and institutional control requirements. Midday sessions will focus on developing defined user scenarios across the three databases and sketching a preliminary technical architecture for cross-schema mediation. By the conclusion of the event, the group will produce a draft governance charter outline, a prioritized set of use cases, and an initial architecture diagram to guide prototype development.

Following the in-person convening, a broader virtual stakeholder event will extend consultation to external copy holders, academic data librarians, repository managers, and additional social scientists. This session will present the draft framework, prioritized use cases, and architecture concept emerging from the core event. Participants will provide additional feedback through polling and moderated discussion. The purpose of this broader consultation is to test whether the identified needs and proposed framework resonate beyond the immediate research community and to assess the potential for national adaptation.

In addition to these convenings, the project team will conduct targeted interviews with researchers, librarians, and IT administrators not represented in the events. These interviews will explore documentation gaps, workflow barriers, evaluation expectations, and institutional readiness for AI-mediated querying. Interview data will complement findings from the events.

All data collected during Phase 2 will be systematically analyzed and synthesized. The team will systematically analyze findings and refine governance and technical specifications accordingly. Outputs from this phase will include a consolidated needs assessment document and finalized architecture requirements. These materials will form the foundation for the proof-of-concept developed in Phase 3 and will inform the public-facing planning report produced in Phase 4.

Phase 3: Proof-of-Concept Design and Technical Feasibility

Phase 3 translates the requirements and governance constraints defined in Phases 1 and 2 into a working proof-of-concept (POC). The goal is to demonstrate that AI-enabled natural language access across three heterogeneous databases is technically feasible within the stewardship boundaries established by stakeholders.

The phase begins with a focused technical review of current approaches to natural language database querying, specifically methods that allow users to ask questions in plain English and receive structured database results, without needing to know the underlying database structure or query language. The review will identify methods best suited to historically layered research databases, with particular attention to reliability, transparency, and error handling. This will inform selection of the technical approach used in the POC.

The team will then address the challenge of working across three databases that were built independently, use different structures, and define variables differently. Rather than merging or restructuring the databases, which would compromise their integrity, the project will create a documented mediation layer that maps comparable concepts across collections, such as governance units, resource systems, and institutional rules. This mapping, developed directly from the priority use cases identified in Phase 2, allows the system to treat the three databases as a coherent set for defined research questions.

Building on this foundation, the team will develop the natural language querying framework. A researcher will be able to pose a question in plain English; the system will dynamically identify the relevant tables and variables across the three databases, generate an appropriate query, and then verify its own output through an iterative self-correction process, catching and resolving errors before results are returned. Governance checkpoints are embedded throughout, ensuring that every query can be inspected, audited, and traced back to the original data structures. This approach is designed to be scalable to large, complex databases without degrading accuracy or transparency.

Internal testing will use predefined research scenarios drawn from the priority use cases. The POC will be evaluated for correctness of data retrieval, consistency across databases, transparency of query logs, and alignment with researcher intent, with domain experts reviewing outputs for semantic accuracy. Testing will occur iteratively, with findings documented in a technical evaluation report. The output of Phase 3 will be a functional but deliberately limited POC prototype, accompanied by a detailed specification of constraints, risks, and requirements for future scaling.

Phase 4: Testing and Dissemination

The concluding Phase 4 will focus on structured evaluation, POC refinement, and dissemination of planning outcomes. Following internal testing in Phase 3, the team will make final revisions to

the POC based on documented technical findings and governance requirements. Revisions will be limited to improving system stability, documentation clarity, and transparency of query generation. The objective is not to expand functionality, but to ensure that the defined use cases operate consistently and that system behavior is fully documented.

The revised POC will then be evaluated by three graduate students representing core user communities in institutional economics, anthropology, and political science. Graduate students are well positioned as evaluators for this planning phase: they are active researchers with domain knowledge and genuine curiosity about the databases, but without the insider familiarity that might mask usability barriers, making them effective proxies for the broader researcher population the system is ultimately designed to serve.

Evaluation will use predefined research scenarios derived from the priority use cases identified in Phase 2. Users will test the system by submitting natural language queries and assessing whether outputs align with research intent, governance expectations, and documented data structures. Evaluation criteria will include correctness of data retrieval, alignment with user intent, cross-database consistency, transparency of query logs, and perceived usability. Structured feedback instruments will document strengths, limitations, and risks.

A virtual event will convene a broader national audience to review and discuss the POC results. This session will present the architecture, governance framework, evaluation findings, and identified constraints. Participants, including academic data librarians, repository managers, external database stewards, and additional social scientists, will discuss results and help identify gaps and limitations. The purpose of this event is to test the broader applicability of the planning framework and to identify potential partners for a future implementation phase.

To disseminate the results of our work, the project team will publish technical documentation describing the architecture, schema mediation approach, governance framework, and evaluation results. Code associated with the limited POC will be made available in a public repository with appropriate documentation and licensing. A public-facing white paper will synthesize stakeholder findings, governance principles, and planning recommendations in a format accessible to libraries and research centers beyond the immediate project community. Webinars will be hosted through IU Libraries and related professional networks to share findings with research data librarians and domain scholars. A dedicated project webpage will centralize access to documentation, recordings, and planning materials.

The final consolidated planning report will outline a concrete pathway for scaling this approach to additional social science databases and partner institutions nationally. By documenting technical feasibility, governance requirements, stakeholder validation, and replication criteria, this project provides the foundation for a subsequent multi-institutional implementation initiative, contributing to the broader goal of making the nation's accumulated social science knowledge accessible to researchers, educators, and civic learners across the country.

Team Structure and Expertise

The project will be directed by Inna Kouper, Associate Scientist at the IU Luddy School of Informatics, Computing and Engineering and Senior Research Fellow at the Ostrom Workshop. Kouper's dual appointment uniquely positions her to lead this project: she brings published

research on data practices, experience managing federally funded research data infrastructure, prior hands-on work with the IFRI, NIIS, and CPR databases, and expertise in AI and computational methods in research contexts.

The project co-Director, Emily Castle, Librarian at the Ostrom Workshop, will oversee stewardship and stakeholder coordination. Castle brings direct operational knowledge of the three databases, longstanding relationships with database stewards and researchers across partner institutions, and expertise in repository management and research support services. Her library perspective is central to ensuring that the governance framework and technical architecture reflect established library values and practices.

Technical development will be led by Dr. Oleksandr Leykin, whose training in computer science and extensive experience in relational databases, natural language processing, and applied machine learning systems directly address the technical challenges of AI-mediated querying across heterogeneous schemas.

The project is institutionally supported by the Ostrom Workshop and IU Libraries Research Data Services (see letters of support), which will contribute expertise in metadata standards, repository integration, and data governance, ensuring that project outputs are grounded in current library practice and positioned for integration into existing institutional infrastructure.

Project Results

This project will produce five interconnected results:

- A governance and coordination framework for responsible AI mediation of research databases
- A technical architecture specification for natural language access across heterogeneous collections
- A functional proof-of-concept demonstrating AI-enabled access across three historically significant social science databases
- A structured needs assessment documenting user requirements, stewardship constraints, and feasibility findings; and
- A publicly available planning framework adaptable by other libraries and research data stewards nationally.

These results directly address a gap that no existing tool or framework currently fills: a replicable, library-governed model for making complex, heterogeneous research databases accessible to researchers, educators, and civic learners without requiring specialized technical expertise. By demonstrating feasibility, documenting governance requirements, and validating a bounded AI mediation approach against defined research scenarios, this project establishes a model that research libraries across the country can adapt to their own collections and institutional contexts.

The outputs are designed from the outset for adaptability and national reach. The governance charter will articulate coordination principles applicable to other distributed research collections. The architecture documentation will describe modular components, covering how database structures are interpreted, queries are generated and validated, and all steps are logged. The needs assessment and evaluation findings will be published in a format that allows other libraries

and research centers to reproduce the planning process, including stakeholder engagement models and feasibility criteria. Code associated with the proof-of-concept will be released with documentation sufficient for adaptation by other institutions.

Dissemination will reach research libraries, specialized data stewards, and social science communities through multiple channels. A public-facing white paper will synthesize stakeholder findings, governance principles, and planning recommendations in an accessible format for libraries and research centers beyond the immediate project community. Project findings will be shared through webinars hosted in collaboration with IU Libraries and relevant professional networks, presentations at national conferences focused on libraries, archives, and research data management, and targeted outreach to research data services consortia. A project website will centralize documentation, architecture diagrams, governance materials, evaluation results, and access to code repositories.

Sustainability is built into both the institutional partnerships and the design of deliverables. The governance framework and planning documentation will be hosted and maintained through Indiana University infrastructure, integrated into ongoing operations of IU Libraries and the Ostrom Workshop. Relationships established among specialized data stewards, research data librarians, and technical developers will continue through existing institutional and professional networks. Most importantly, the planning report produced in Phase 4 will provide a concrete, documented pathway, including technical specifications, governance requirements, and replication criteria, for subsequent initiatives to scale AI-enabled access across additional social science collections nationally through IMLS National Implementation grants, NSF data infrastructure programs, and private foundations supporting digital scholarship.

At a moment when libraries are navigating both the pressures and the possibilities of AI adoption, this project demonstrates that responsible integration is achievable through partnerships, stakeholder participation, and technical experimentation. By establishing a replicable, governance-grounded model for AI-enabled access to complex research data, this project advances libraries' national role as trusted intermediaries in the knowledge infrastructure of American civic and scholarly life.

Project Title: AI-Enabled Access and Stewardship for Heterogeneous Social Science Databases in Libraries and Archives								
Schedule of Completion								
	Year 1 (Sept 2026 - July 2027)				Year 2 (Aug 2027 - July 2028)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1: Partnership Structuring and Governance.								
Outreach to database stewards & copy holders outside IU	X							
Regular project meetings and discussions	X	X	X	X	X	X	X	X
Landscape analysis of database copies	X							
Collaboration framework and governance charter	X							
Review of AI approaches to database access and use	X	X						
2: Needs Assessment.								
Core stakeholder event (in-person)		X						
Broader stakeholder event (virtual)		X						
Additional stakeholder interviews (researchers, librarians, IT)		X	X					
Data analysis and triangulation			X	X				
Data organization and documentation	X	X	X	X	X	X	X	X
3: Proof-of-Concept.								
Best practices review (AI-to-database mediation)			X					
Architecture design			X	X				
Cross-database alignment				X	X			
Natural language querying framework development				X	X			
Internal proof-of-concept testing and quality assurance					X	X		
4: Testing and Dissemination.								
Final revisions of the proof-of-concept approach					X	X		
Evaluation by core users (testers)						X	X	
Virtual event (POC evaluation + national discussion)								
Dissemination (webinars, code, website)							X	X
Final project documentation and wrap-up								X