



LAURA BUSH 21st CENTURY LIBRARIAN PROGRAM

Sample Application RE-260375-OLS-26

Project Category: Applied Research

Indiana University, Luddy School of Informatics, Computing,
and Engineering

Amount awarded by IMLS: \$671,369

Amount of cost share: \$0

Indiana University Bloomington will partner with several institutions to address a critical gap in library and archival education and workforce preparation, identified through the professional literature and practitioner consultation. Artificial Intelligence (AI) systems leave behind artifacts, and institutions are already accumulating these data without guidance for assessing their evidential value, risks, or stewardship implications. This applied research project will conduct a mixed-methods study to determine what library and archival professionals must understand about AI data to make responsible stewardship decisions. The project will lead to a conceptual framework, pilot a graduate curriculum module across four partner institutions, and adapt it for national use.

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.

DECODE: Developing Competencies for AI Data Evaluation in Library and Archival Education

INTRODUCTION

Indiana University Bloomington (IUB) requests \$671,209 in Institute of Museum and Library Services (IMLS) funds for DECODE: Developing Competencies for **AI Data** Evaluation in Library and Archival Education, a three-year Applied Research project addressing a foundational gap in LIS professional knowledge about **AI data**. The field lacks frameworks for understanding what **AI data** are, what risks they pose, and what practitioners would need to know before they can make stewardship decisions about **AI data**. Libraries and archives are already encountering **AI data**, such as training datasets, model files, learned parameters, configuration files, and system documentation, without professional guidance for evaluating their risks, evidential value, or long-term stewardship implications. DECODE develops the foundational knowledge practitioners need to make those decisions responsibly. DECODE is led by IUB in collaboration with the Northeast Document Conservation Center (NEDCC), pilot site partners at the University of Maryland, College Park (UMD), the University of Pittsburgh (Pitt), and North Carolina Central University (NCCU), and project team members Dr. Tao Dong (Senior Staff User Experience Researcher, Core Developer Division, Google) and Dr. Sarah Buchanan (Convener, Archival/Preservation Education (ArchiPrEd) Special Interest Group, Association for Library and Information Science Education (ALISE)). DECODE's research questions are: (1) What are the essential types, characteristics, and risk profiles of **AI data** that LIS practitioners need to understand to make informed stewardship decisions? (2) What knowledge gaps exist among current practitioners and students **about the data that drive AI**? (3) How can foundational knowledge about **AI data** be effectively integrated into LIS graduate education? Deliverables include a conceptual framework translating AI technical knowledge for LIS audiences, a competency framework identifying practitioner learning needs, a tested graduate curriculum module, an on-demand professional development course through NEDCC's Training Hub, and a peer-reviewed journal article. Consistent with Executive Orders 14277 and 14278, and in direct response to OLS's stated priority to support projects that creatively and effectively use artificial intelligence to improve academic practices, this project strengthens AI-related education and workforce readiness by building the foundational knowledge practitioners need to engage responsibly with AI-related materials. Project deliverables are designed for delivery across multiple educational models, including ALA-accredited Master of Library Science (MLS) programs, continuing education, and, through their openly licensed and modular design, potential adaptation to alternative credentialing pathways by programs and organizations best positioned to implement them. This project aligns with LB21 Program Goal 2 (Train and retain current library and archives professionals, faculty, and staff) and Objective 2.1 (Develop professional development and training programs to enable the library and archival workforce to meet the needs of their communities).

PROJECT JUSTIFICATION

Statement of Need

Archivists and librarians are the professionals society has designated to preserve the record of consequential decisions. AI systems are now making consequential decisions at scale across research, government, healthcare, education, and cultural life. Those systems leave behind artifacts, including AI data such as training data, model weights, configuration files, evaluation logs, and prompt histories. Whether those artifacts are records, what risks they carry, and what practitioners would need to know before any stewardship decision could responsibly be made are questions that do not yet have clear answers in the professional literature. AI data exist today, so the profession can anticipate future stewardship challenges. **Consider a university archivist who receives a transfer from a research center that closed after a decade of using AI systems to analyze oral history collections, flag records for appraisal, and generate descriptive metadata across tens of thousands of items. She has no framework for these questions. Neither does her professional literature.** DECODE addresses this gap, which has immediate implications for appraisal, access, and institutional risk.

Artificial intelligence systems increasingly shape information access and public decision-making (Crawford, 2021). Regulatory frameworks such as the EU AI Act (2024) presume that AI systems and their documentation will be available for retrospective examination. Archivists and digital preservation professionals already encounter records generated by or dependent upon automated decision-making systems and must determine how existing appraisal and stewardship frameworks apply. Unlike purely technical infrastructure managers, archivists are responsible for long-term evidential stewardship, appraisal, and access decisions that require understanding the informational objects themselves. This need has been confirmed through recent professional community activity. The Digital Library Federation's Born-Digital Access Working Group received a 2024 Society of American Archivists Foundation grant specifically to investigate AI adoption patterns and barriers among American archival workers, reflecting the profession's own recognition that foundational knowledge about AI engagement remains underdeveloped (Digital Library Federation, 2024). The Executive Director of the Digital Preservation Coalition has similarly called for urgent action to establish frameworks for assessing AI-related

content and build momentum toward AI preservation at a systemic level, characterizing the challenge as a matter of public accountability (Kilbride, 2024).

Consider a processing archivist at a regional historical society completing appraisal of a digital transfer from a dissolved nonprofit. Among the files are directories labeled models/, checkpoints/, and training logs—AI data produced during model development and use. The organization relied on these systems to analyze client records and flag cases for advocacy over five years. She must decide within the week whether to ingest, quarantine, or destroy these files. She knows the underlying records are sensitive. She does not know whether the model weights could allow someone to reconstruct them. She cannot find a professional standard, a continuing education resource, or a colleague who knows the answer. She makes her best guess and moves on.

The fundamental issue is that the cultural heritage professions lack frameworks for understanding what **AI data** are, how they are acquired, organized, managed, modified, tested, trained, or improved. As Padilla (2025) argues, AI preservation "remains to a large extent unaddressed," with basic questions about what should be preserved and how documentation practices should differ from current approaches still unanswered. A comprehensive review found "a lack of shared understanding of what AI actually entails" among archival professionals (Colavizza et al., 2022, p. 4). The scale of this gap has now been documented empirically. A nationally distributed survey of 760 U.S. academic library employees found that readiness for AI implementation was low across the workforce, with participants reporting modest self-rated AI understanding and significant deficits in the technical dimensions of AI literacy (Lo, 2024). A 2025 global study of more than 2,000 librarians found that despite 67% exploring or implementing AI tools, only 33% had moved beyond initial evaluation, and U.S. librarians ranked last in confidence, with just 7% expressing optimism about AI's benefits over the next five years (Clarivate, 2025).

Emerging scholarship in AI security further complicates the picture: some AI data artifacts may be dangerous to open, as trained model weights can enable extraction of private training data (Carlini et al., 2021). Others may be dangerous to keep, as poisoned datasets and backdoored models represent persistent institutional risks if preserved without systematic auditing (Chen et al., 2017). The profession has not begun to grapple with these distinctions. **Whether librarians and archivists have any role in AI data stewardship remains an open question that cannot be answered until the profession develops foundational knowledge about AI data.** Regardless of the ultimate jurisdictional determination, practitioners require foundational literacy in **AI data** structures to make informed professional and institutional decisions. This need is made more urgent by structural pressures facing LIS professional education: program closures, declining enrollments, and growing competition from alternative credentialing pathways. If LIS programs cannot demonstrate relevance to emerging information challenges like AI, the case for their continued existence weakens. This project produces outputs that function across multiple educational models, whether delivered through traditional graduate programs, continuing education, or micro-credential pathways, so that the knowledge reaches practitioners regardless of how professional education evolves.

Relationship to Existing Models, Scholarship, and Practice

Existing LIS scholarship on AI has developed along two primary tracks, neither of which addresses the question this project investigates. The first track examines AI as a tool for library and archival work: automating description, enhancing discovery, and supporting appraisal decisions. Al-Mutairi's (2025) doctoral research exemplifies this approach, developing a framework for integrating AI into the archival value chain. The second track examines AI as a social and political phenomenon requiring critical analysis (Crawford, 2021). This project occupies an alternative, largely unaddressed position: **AI data** as informational objects that practitioners may be asked to appraise, steward, or make accessible, and for which the profession currently lacks foundational knowledge.

Within digital preservation practice, existing frameworks and standards were developed for records, datasets, and software, not for the composite, dynamic artifacts that constitute AI systems. Padilla (2025) has identified AI preservation as an unaddressed challenge; this project responds directly by investigating what practitioners would need to know before stewardship decisions could responsibly be made. Lynch (2017) similarly argued that traditional archival approaches are inadequate for algorithmic systems but stopped short of specifying what alternative knowledge base practitioners would require. The AI and machine learning (ML) community has produced documentation frameworks, including Model Cards (Mitchell et al., 2019), Datasheets for Datasets (Gebru et al., 2021), and Data Statements (Bender & Friedman, 2018), representing current practice for recording what matters about AI systems. However, these assume technical literacy that LIS professionals do not yet possess, and Raji et al. (2020) demonstrate persistent gaps between documentation aspirations and organizational practice. This project complements rather than duplicates that body of work by translating its core concepts into terms accessible to LIS audiences and identifying which elements are most relevant to stewardship assessment.

Beneficiaries

Primary beneficiaries include practicing archivists, librarians, LIS educators, and graduate students, who will gain research-grounded resources and foundational knowledge about **AI data** to inform their professional practice. Pilot testing at UMD,

Pitt, NCCU, and IUB ensures that the curriculum is validated across distinct institutional missions and student populations, including research-intensive institutions and the only historically Black university with an ALA-accredited LIS program. Adaptation into an on-demand course with enrollment scholarships extends access to practitioners nationwide, including those without institutional travel or professional development budgets. Secondary beneficiaries include the AI/ML community, which may benefit from engagement with long-term stewardship expertise in the LIS community, and institutions navigating records management obligations for AI-generated outputs.

The need these beneficiaries face is not hypothetical. The problem has already arrived. A special collections librarian has spent the past year using an AI-assisted description tool to process several thousand digitized items from a civil rights collection. The descriptions are consistent, her colleagues approve, and the backlog is moving. A faculty member asks two questions before assigning the collection in a graduate seminar: what data was this tool trained on, and could this collection itself have been part of that training? She contacts the vendor. The vendor cites proprietary constraints. She has no framework for assessing what she would need to know about this tool to provide honest assurances of provenance, reproducibility, or freedom from embedded bias. She cannot document the tool in the finding aid in any way that would satisfy a researcher who needed to know. The knowledge DECODE will produce is not preparation for a future challenge. It is already overdue.

PROJECT WORK PLAN

Research Questions

RQ1: What are the essential types, characteristics, and risk profiles of **AI data** that LIS practitioners need to understand to make informed stewardship decisions?

RQ2: What knowledge gaps exist among current practitioners and students about **the data that drive AI**?

RQ3: How can foundational knowledge about **AI data** be effectively integrated into LIS graduate education?

Methods

The project employs a four-phase mixed-methods design progressing from technical knowledge gathering to practitioner needs assessment to curriculum development and testing to practitioner education delivery. A mixed-methods design is appropriate for this inquiry because the research questions require both the depth of qualitative exploration to surface tacit practitioner knowledge and the breadth of survey data to assess curriculum coverage nationally. Table 1 summarizes the phase structure, timelines, and deliverables.

Table 1. DECODE Phase Structure

Phase	Months	RQ	Deliverable	Key Activities
1	1-8	RQ1	Conceptual Framework	Systematic review; AI practitioner interviews; documentation artifact analysis
2	6-14	RQ2	Competency Framework	LIS practitioner interviews; graduate student focus groups; LIS faculty survey
3	12-24	RQ3	Curriculum Module; Journal Article	Module development; 4-site pilot testing with pre/post assessment; manuscript
4	25-36	All	On-Demand Course	Course production with NEDCC; summative assessment design; scholarship launch; conference presentation

Phase 1: Understanding AI Data (Months 1 to 8) addresses RQ1 through three complementary methods. (a) *Systematic review* of artificial intelligence and machine learning (AI/ML) literature on system components, data practices, and documentation frameworks, synthesizing Sculley et al. (2015), Sambasivan et al. (2021), Paullada et al. (2021), and related scholarship. These works provide the technical grounding for Phase 1's synthesis and are situated more fully within the project's Theoretical and Conceptual Framework described below. Sources will be coded deductively against AI lifecycle categories (data acquisition, model training, deployment, and documentation) and inductively to surface concepts without existing LIS equivalents, ensuring the Conceptual Framework reflects both established AI/ML vocabulary and emergent practitioner-relevant distinctions. (b) *Semi-structured interviews* (approximately 60 minutes each, conducted via Zoom, recorded with participant consent, and transcribed verbatim) with 15 to 20 AI practitioners (researchers, engineers, data scientists) recruited through multiple channels, including AI research communities at IUB, professional conference networks, and supplemental outreach through Dr. Dong's professional contacts, LinkedIn network, and in-person engagement with AI practitioners at major venues such as CHI 2026 and AIWare, where study information may be shared

with prospective participants. Pre-project outreach has already identified prospective participants whose work spans trustworthy AI, AI security, ML systems, and real-world deployment contexts, and who have expressed willingness to participate; one prospective participant has additionally offered to facilitate introductions to colleagues conducting related research, demonstrating that the Phase 1 recruitment pipeline is substantively developed prior to the project start date. Incorporating AI security perspectives ensures that risk is examined in terms of how systems can fail, be manipulated, or expose sensitive information in practice, complementing policy-oriented discussions and grounding preservation risk categories in real-world system behavior. Interview questions will probe how **AI data** are acquired, organized, managed, modified, tested, trained, and improved in practice, and will specifically investigate which categories of **AI data** practitioners consider unsuitable for open preservation or long-term archival retention, including artifacts that may carry biometric, proprietary, or security-sensitive information. A randomly selected subset of transcripts will be returned to participants for member-checking prior to analysis to support validity. (c) *Document analysis* of existing AI documentation artifacts (model cards, datasheets, data statements) to identify what is documented, what is omitted, and what would be needed for long-term stewardship assessment. The Phase 1 IRB application, covering interview procedures and consent protocols, will be submitted at project launch with approval anticipated prior to the start of data collection. IUB maintains an active Federal-Wide Assurance. Phase 1 deliverable: A Conceptual Framework.

Phase 2: Understanding Practitioner Knowledge Gaps (Months 6 to 14) addresses RQ2 through three methods. (a) *Semi-structured interviews* (approximately 60 minutes each, via Zoom, recorded with consent, transcribed verbatim) with 20 to 25 practicing archivists, librarians, and digital preservation specialists recruited through professional listservs and networks of the Midwest Archives Conference (MAC), Society of American Archivists (SAA), the Digital Library Federation (DLF), and the International Conference on Digital Preservation (iPRES) communities. Participants will be purposively sampled to represent variation in institutional type (academic, government, special, public), career stage, and geographic region to ensure broad applicability of findings. Interview questions will probe participants' existing conceptualizations of AI, their encounters with AI systems and **AI data** in their collections or workflows, and their perceived readiness to make stewardship decisions about **AI data**. (b) *Focus groups* with LIS graduate students (3 to 4 groups of 6 to 8 students) recruited at IUB, conducted by the Project Director via Zoom, recorded with consent, and transcribed verbatim. Focus group questions will probe students' existing conceptualizations of **AI data**, their understanding of how AI systems are built and documented, and where they perceive gaps in their own preparation for stewardship decisions involving **AI data**, providing the empirical grounding for curriculum module development in Phase 3. (c) *Online survey* of LIS faculty distributed through professional LIS education networks, including the ALISE ArchiPrEd Special Interest Group (ArchiPrEd SIG) and related faculty communities, to assess current curriculum coverage, perceived gaps, and faculty readiness to integrate new instructional materials. The instrument will be pilot tested with a small sample of LIS faculty recruited through these professional networks prior to full distribution to ensure clarity and construct validity. Prior to Phase 2 data collection, an IRB amendment will be submitted covering the interview and focus group protocols and the faculty survey instrument, which will be developed during Phase 2 and reviewed by IRB before distribution.

Phase 3: Curriculum Development and Testing (Months 12 to 24) addresses RQ3 through: (a) *Curriculum module development* beginning at Month 12. The module will be organized around three learning objectives, each grounded directly in the Competency Framework: (1) identifying and distinguishing major categories of **AI data** and their stewardship implications; (2) interpreting existing AI documentation frameworks in terms relevant to archival appraisal; and (3) recognizing which categories of **AI data** may pose safety, privacy, or security concerns if preserved or made openly available. Curriculum module development will begin at Month 12 in parallel with finalization of the Competency Framework and will be completed following the Framework's completion at Month 14, ensuring the module is empirically grounded rather than assumption driven. Dr. Buchanan (ArchiPrEd SIG Convener) will coordinate faculty feedback on the curriculum module through ArchiPrEd SIG programming prior to pilot testing. (b) *Pilot testing* at IUB, UMD, Pitt, and NCCU during Spring 2028, with pre/post assessment measuring changes across the three learning objective constructs. Effect sizes will be calculated to assess the practical significance of learning gains across all sites. (c) Revision based on pilot feedback from the instructors. (d) Manuscript preparation synthesizing empirical findings into recommendations for the profession. The pre/post assessment protocol and instructor feedback procedures will be covered under the Phase 2 IRB amendment, submitted prior to pilot testing, and are anticipated to qualify for exempt status as research conducted in established educational settings involving normal educational practices. Phase 3 deliverables: Graduate curriculum module with instructor guide; Journal article manuscript.

Phase 4: Practitioner Education Delivery (Months 25 to 36) includes: (a) Collaboration with NEDCC to determine optimal format, scope, and delivery structure informed by pilot feedback; (b) Course production by NEDCC, adapting the curriculum module for self-paced professional development (anticipated duration approximately 3–4 hours of modular instructional content), including a summative learning assessment aligned to the Competency Framework: the PD will provide learning objectives and guidance on key concepts drawn from Phase 2 findings; NEDCC's instructional designers will produce the question bank, distractors, and immediate learner feedback; and the PD will review and approve the final

instrument prior to launch; (c) Launch with 45 enrollment scholarships, awarded through an application process administered by NEDCC with criteria developed collaboratively by NEDCC and the PD; NEDCC will also promote the course and scholarship opportunity through its established communication channels, including its national newsletter and professional mailing lists; (d) Digital Directions conference presentation with 28 registration scholarships; (e) Development of a maintenance and updating plan, including a content review schedule. Scholarship recipient feedback surveys will be covered under a Phase 4 IRB amendment submitted prior to survey distribution and are anticipated to qualify for exempt status. Phase 4 deliverables: on-demand course; Digital Directions presentation; Practitioner Brief #3. The sequential phase structure ensures each phase produces empirically grounded outputs informing subsequent activities. Deliberate phase overlaps mitigate sequencing risk: Phase 2 begins at Month 6 while Phase 1 continues; curriculum development begins at Month 12 using preliminary Phase 2 findings as scaffolding before the Competency Framework is complete at Month 14. Formative evaluation at Months 14 to 15 and biannual advisory board meetings provide structured checkpoints throughout. The primary sequencing risks are IRB delays and participant recruitment shortfalls. These are mitigated by submitting the IRB application in Month 1, leveraging pre-established recruitment pathways confirmed before the project start date, and building redundant recruitment channels across multiple professional venues and networks in both phases. Should participant recruitment fall short of targets in either phase, the advisory board's practitioner and faculty networks provide an additional recruitment channel activated at the Month 6 meeting.

Theoretical and Conceptual Framework

This project addresses the inverse of Al-Mutairi's (2025) work: rather than examining how AI can assist archives, it investigates what archivists need to understand about **AI data** to make informed stewardship decisions. This project is grounded in scholarship recognizing AI as a distinct class of socio-technical object requiring new conceptual frameworks. Real-world ML systems are dominated not by model code but by data infrastructures, including collection pipelines, verification processes, and monitoring systems (Sculley et al., 2015), while data work itself is often undervalued and inconsistently documented (Sambasivan et al., 2021; Paullada et al., 2021). Here, "**AI data**" refer to the artifacts produced, transformed, and relied upon across the AI lifecycle: training datasets, learned parameters, configuration files, evaluation benchmarks, documentation, and adversarial artifacts whose preservation raises distinct safety and governance concerns. **The project distinguishes AI systems as operational infrastructures from AI data as informational artifacts that may constitute the primary objects of archival assessment.** Refining this definition through empirical investigation is a core objective of Phase 1.

Burrell (2016) identifies three forms of opacity characterizing ML systems: intentional corporate secrecy, technical illiteracy among non-specialists, and intrinsic opacity that renders systems opaque even to experts. Selbst et al.'s (2019) "Portability Trap" cautions against assuming that frameworks developed for one context transfer directly to another, suggesting that existing digital preservation frameworks cannot be assumed to apply to AI without significant conceptual work. Abbott's (1988) theory of professional jurisdiction provides a framework for understanding whether and how LIS might claim responsibility for AI stewardship: professions secure jurisdiction over problem domains by developing abstract knowledge capable of redefining problems in profession-specific terms. Whether LIS can develop such knowledge for AI, and what form it would take, is precisely what this project investigates empirically.

Plans for Analysis

Each data source employs analysis methods appropriate to its form, managed through NVivo 15 qualitative data analysis software, which serves as the central platform for organizing, coding, querying, and documenting all qualitative and mixed-methods data across Phases 1 and 2. NVivo's mixed-methods functionality supports integration of qualitative findings with Phase 2 survey data and Phase 3 pre/post assessment results, enabling systematic cross-dataset comparison; all codebooks, memos, and decision trails will be maintained in NVivo to support transparency and replicability.

The Phase 1 systematic review will use a structured synthesis protocol, with sources coded deductively against AI lifecycle categories and inductively to surface concepts without existing LIS equivalents. Existing AI documentation artifacts will undergo directed content analysis, coding first against known documentation categories and then open-coding for stewardship-relevant elements. Phase 1 interview data will be analyzed using thematic analysis (Braun & Clarke, 2006), allowing both deductive coding against the emerging conceptual framework and inductive identification of practitioner knowledge not mapping to existing categories. All qualitative coding will employ inter-rater reliability procedures: the Project Director and graduate research assistant will independently code a randomly selected 20% subsample of each data type, with inter-rater reliability calculated using Cohen's Kappa at a threshold of $k \geq 0.80$ (Landis & Koch, 1977). Where Kappa falls below threshold, coders will reconcile disagreements through structured discussion and recalculate before proceeding.

Phase 2 qualitative data (LIS practitioner interviews and LIS student focus groups) will be analyzed using thematic analysis in NVivo, consistent with the procedures described above. Systematic cross-dataset comparison of AI practitioner perspectives (Phase 1) and LIS practitioner perspectives (Phase 2) will enable identification of where knowledge gaps are

most pronounced and where practitioner vocabularies diverge most sharply. The inter-rater reliability procedures specified for Phase 1 will apply across all qualitative datasets. Faculty survey data will be analyzed using descriptive statistics for closed-ended items and thematic coding for open-ended responses, triangulated against qualitative findings to assess convergence. Phase 3 pre/post pilot data will be analyzed using paired comparisons, with effect sizes calculated to assess the practical significance of learning gains across all three pilot sites.

Project performance will be monitored using the LB21 program performance measures of effectiveness, efficiency, quality, and timeliness, ensuring that project activities and outputs are assessed systematically throughout the grant period.

Generalizability is ensured through deliberate sampling heterogeneity at every phase. AI practitioner interviews will span academic, industry, and government research contexts. LIS practitioner interviews will include archivists, librarians, and digital preservation specialists across institutional types, career stages, and regions. Faculty survey distribution through professional LIS education networks, including the ArchiPrEd SIG, targets broad national reach across the LIS educator community. Pilot testing across IUB, UMD, Pitt, and NCCU ensures curriculum validation across distinct institutional missions, student populations, and resource contexts.

Project Team and Stakeholder Engagement

Project Director (PD): Devan Ray Donaldson, Ph.D., Associate Professor and Director of the MLS Program, IUB, who also directs the Archives and Records Management specialization, providing a direct pipeline for integrating project findings into the graduate curriculum. Dr. Donaldson's peer-reviewed research demonstrates sustained methodological expertise directly applicable to this project: his focus group studies of data sharing practices, interview-based research on how data managers conceptualize the Open Archival Information System (OAIS) Reference Model and the FAIR Guiding Principles for Scientific Data Management and Stewardship, and content analysis of how institutions represent trustworthiness employ precisely the methods used in Phases 1 and 2. His prior research developing the Trust in Archives-Trust in Digital Archival Content (TIA-TDAC) Framework (Donaldson, 2019) and the Digitized Archival Document Trustworthiness Scale (DADTS) (Donaldson, 2016) for assessing trustworthiness in digital archival contexts establishes the intellectual foundation from which DECODE's AI preservation trustworthiness questions naturally emerge, making him uniquely positioned to lead this inquiry. He is a 2017-2018 Indiana University Trustees Teaching Award recipient and 2024 Luddy Faculty Recognition Award recipient. As PD, Dr. Donaldson will provide overall project leadership and management, design and oversee all data collection protocols, lead qualitative analysis across Phases 1 and 2, direct curriculum development in Phase 3, and coordinate all dissemination and evaluation activities throughout the project period.

AI Industry Consultant: Tao Dong, Ph.D., Senior Staff User Experience Researcher, Core Developer Division, Google (Human-Centered AI in software development), participating in a personal capacity. Dr. Dong brings direct expertise in how AI systems are built, deployed, and documented in industry contexts. He will advise on Phase 1 interview protocol development and conceptual framing to ensure alignment with contemporary AI industry practice; support recruitment efforts through dissemination to his professional contacts, LinkedIn network, and professional engagement at relevant AI venues; provide guidance on identifying appropriate participant criteria to ensure representation across AI domains; review and provide substantive feedback on the Conceptual Framework for technical accuracy and industry relevance; and contribute to ALISE 2027 dissemination activities.

ArchiPrEd SIG Partner: Sarah Buchanan, Ph.D., Associate Professor and Emphasis Leader of Archival Studies, iSchool (School of Information Science & Learning Technologies), University of Missouri, brings research and teaching expertise in archival practice, appraisal, data and records management, and computational approaches to archival education. Her survey research on tasks and competencies to inform records management education (Buchanan et al., 2017) is directly relevant to the Competency Framework development in Phase 2, and her peer-reviewed work on computational thinking integration into archival educators' networked instruction (Buchanan et al., 2022) speaks directly to the curriculum development activities in Phase 3. As a collaborative research member of two IMLS-funded projects, the CT Piloting Network (2020 to 2022) and the TALENT Network (2022 to 2025), she brings firsthand experience with IMLS-funded curriculum innovation in LIS contexts. As Convener of the Archival/Preservation Education (ArchiPrEd) Special Interest Group of ALISE since 2018, Dr. Buchanan leads a SIG of 75+ members as of 2025, providing direct access to the national community of LIS faculty engaged in archival and preservation education, the primary audience for curriculum adoption. The SIG's established governance structure and conference presence ensure this channel persists as an institutional resource independent of any individual convener. During Phases 2 and 3, Dr. Buchanan will advise on curriculum development in her individual capacity, support the pilot test and distribution of the LIS faculty online survey, coordinate ArchiPrEd SIG faculty review of the curriculum module prior to its classroom-based pilot testing, coordinate and facilitate the DECODE presentation at the ArchiPrEd SIG session during ALISE 2027, and support ALISE 2028 dissemination activities.

Practitioner Education Partner: Northeast Document Conservation Center (NEDCC). Founded in 1973, NEDCC is the nation's first nonprofit center devoted to the conservation of cultural heritage materials and has a long history of continuing education for preservation professionals. Its Training Hub, which offers 70 live and on-demand professional development

courses each year for over 1,500 registrants, provides a proven delivery platform with demonstrated reach and established infrastructure for course production and learner management. The Fundamentals of AV Preservation web course, originally developed with grant support, remains active in the Training Hub catalog, demonstrating that NEDCC's cost-recovery model successfully sustains courses beyond their initial funding period. Its Digital Directions conference, now in its 19th year, has reached 2,070 practitioners across its history. Training Hub learners have represented organizations in all 50 states, demonstrating national reach. NEDCC will develop and deliver the Phase 4 on-demand course, administer scholarship awards, and sustain the course through its standard cost-recovery catalog for a minimum of one year after the grant period. Advisory board member Ann Marie Willer (Director of Preservation Services, NEDCC) provides institutional continuity between the research and delivery phases.

Advisory Board: A five-member advisory board will provide guidance at key milestones (Months 6, 12, 18, 24, 30, 36). Written progress reports summarizing phase deliverables and emerging findings will be circulated in advance of each meeting to enable substantive, milestone-specific feedback rather than general discussion. The Month 36 meeting is specifically structured as a board review of the summative evaluation report prior to finalization, ensuring community oversight through the close of the project. **Derek T. Mosley** (80th President, Society of American Archivists (SAA); Archives Division Manager, Auburn Avenue Research Library on African American Culture and History, Atlanta, GA) brings the perspective of the profession's leading national figure and a senior archivist actively navigating AI-generated records in a culturally significant special collections context. As SAA President, Mr. Mosley provides a direct channel for communicating findings to the broadest constituency of archival practitioners nationally. **Holly Smith** (College Archivist, The Women's Research and Resource Center, Spelman College) contributes expertise in community archives and specialized archival collections at Historically Black Colleges and Universities (HBCUs). **Dr. Kimberly Anderson** (Director, Distinctive Collections and Associate Professor, University of Nevada, Reno) brings practitioner and research expertise in archival appraisal and special collections management, including leading active exploration with her staff on the impact of AI use and AI-generated output on appraisal practices and research services. **Ann Marie Willer** (Director of Preservation Services, NEDCC) brings digital preservation expertise and experience as principal investigator on grants totaling \$6.6 million. **Thomas Padilla** (Associate Dean for Research and Learning, University of Nebraska-Lincoln Libraries) brings pioneering scholarship on library engagement with computational materials and AI preservation challenges.

Archival Practice Consultant: Dinah Miles (Digital Archivist, IU University Archives) will advise on grounding the Phase 3 curriculum module in contemporary archival workflows and review draft learning objectives, instructional activities, and assessments prior to pilot testing. Drawing on her direct experience applying AI tools and collaborating with AI consultants in an active archival setting, she will ensure the module reflects authentic archival decision-making contexts rather than hypothetical scenarios. She will provide implementation-focused feedback during the revision process to strengthen the module's practical applicability across varied institutional environments. Her contribution will enhance the module's grounding in applied archival practice and increase its relevance for both graduate students and working professionals.

External Evaluator: Paul Conway, Ph.D., Associate Professor Emeritus of Information, University of Michigan School of Information (UMSI), brings extensive expertise in digital preservation, archival quality assessment, and evaluation methodology directly relevant to DECODE's research focus. Over three decades, Dr. Conway has pioneered research on the formative evaluation of archival programs, the summative assessment of large-scale digitization projects, and frameworks for managing machine dependent archival collections. Dr. Conway has served as a reviewer for IMLS National Leadership Grants (1999, 2000, 2011) and the IMLS Laura Bush 21st Century Librarian Program (2011, 2013, 2016, 2017), giving him direct familiarity with IMLS evaluation standards and expectations. His prior collaboration with the Project Director on foundational research in digital archival trustworthiness and his service on the NEDCC Advisory Committee (1994 to 2000) provide continuity across the project's core intellectual and partnership dimensions. Dr. Conway also served as one of several advisors during proposal development. As external evaluator, he will serve at arm's length, with no role in the setup or administration of project activities. In a half-time appointment at UMSI, Dr. Conway remains available for focused external evaluation commitments. He conducts the mid-project formative assessment (Months 14–15) and serves as lead reviewer of the summative evaluation report (Months 33–36).

Graduate Research Assistants (GRAs): A doctoral GRA will be recruited in Year 1 and supported throughout the three-year project. The GRA will contribute to all phases: managing the systematic review database and coding in Phase 1; coordinating interview and focus group scheduling, conducting transcription quality review, and contributing to thematic analysis in Phases 1 and 2; supporting curriculum module drafting and pre/post instrument development in Phase 3; and assisting with course review and evaluation data compilation in Phase 4. The GRA will also assist the PD in preparing and presenting project findings at conferences as part of their professional development. Two MLS hourly students will provide additional research support in Years 1 through 3, assisting with literature searches, document coding, and data organization.

Involving MLS students in active research contributes a direct pedagogical benefit: students gain firsthand experience with the professional knowledge gaps this project addresses.

Pilot Site Partners: Richard Marciano, Ph.D., Professor, College of Information, University of Maryland, College Park; Nora Mattern, Ph.D., Associate Teaching Professor, School of Computing and Information, University of Pittsburgh; and Heather Barnes, Ph.D., Assistant Professor, Library and Information Sciences, North Carolina Central University, will each identify an appropriate LIS course at their respective institutions in which to integrate the curriculum module in Spring 2028. At each site, the module will be implemented, pre- and post-assessment data will be collected, and structured implementation feedback will be provided. Together with IUB, these pilot sites ensure cross-institutional validation across distinct institutional missions and student populations.

Stakeholder Engagement

Stakeholders have been engaged throughout project development. Ann Marie Willer provided input that shaped the project's three-year structure and the design of Phase 4's practitioner education activities, drawing on NEDCC's experience delivering sustained continuing education. Dr. Paul Conway (Associate Professor Emeritus, University of Michigan School of Information) provided feedback that sharpened the project's focus on **AI data**, rather than AI systems broadly, as the primary locus of preservation assessment. Pre-project engagement with AI researchers has strengthened the Phase 1 recruitment pipeline: Dr. Hanqing Guo (Assistant Professor, Departments of Electrical and Computer Engineering and Information and Computer Sciences, University of Hawaii at Manoa), whose research on trustworthy AI, AI security vulnerabilities, backdoor attacks, and voice identification systems maps directly onto Phase 1's investigative focus, has expressed willingness to participate in Phase 1 interviews and offered to facilitate introductions to colleagues conducting related research, demonstrating that the Phase 1 recruitment pipeline is substantively developed prior to the project start date. Partner commitments from NEDCC, pilot site partners, and advisory board members are formalized in letters of participation included as supporting documents. Dr. Buchanan's individual commitment to advise on curriculum development and coordinate ArchiPrEd SIG review is similarly confirmed. Ongoing stakeholder input throughout the project period is structured through biannual advisory board reviews, the dedicated ArchiPrEd SIG feedback session at ALISE 2027, the mid-project formative evaluation by Dr. Conway, and structured pilot site implementation feedback, as described in the Evaluation section.

Evaluation

Formative evaluation will occur through the biannual advisory board meetings described above, with each meeting structured around written progress reports circulated in advance to enable substantive, milestone-specific feedback. A dedicated ArchiPrEd SIG session during ALISE 2027 (November 2027), coordinated through Dr. Buchanan as SIG Convener, will present Phase 1 findings and the Conceptual Framework, featuring the PD, Dr. Tao Dong, and Thomas Padilla, and gather structured feedback from the 75-member SIG community to inform Phase 3 curriculum development. External evaluator Dr. Paul Conway will conduct a mid-project formative assessment (Months 14 to 15), reviewing Phase 1 and early Phase 2 processes and providing written recommendations for Phase 3 refinement. Pilot testing includes pre/post assessment of student learning measured against the three Competency Framework constructs, and instructor implementation feedback collected through structured reflection protocols. **Summative evaluation** in Months 33 to 36 will assess: (a) quality and uptake of deliverables, measured by download counts from IUScholarWorks and OER Commons, citations, and documented adoptions; (b) impact on LIS curriculum, measured by the number of programs that formally integrate module content, tracked through follow-up with pilot site instructors and ALISE network contacts; (c) influence on professional discourse, assessed through citation tracking and evidence of the journal article shaping subsequent scholarship; (d) reach of practitioner-facing dissemination, measured by scholarship recipient feedback surveys, webinar attendance records, practitioner brief download counts, and project website analytics; and (e) on-demand course enrollment, completion rates, and aggregate learner performance on the summative assessment, reported by NEDCC as part of standard Training Hub analytics.

Indicators of Success. Project success will be measured through learning outcomes and evidence of field engagement. Meaningful student learning gains will be demonstrated through statistically significant pre/post differences ($p < .05$) and small-to-moderate effect sizes (Cohen's $d \geq 0.30$), with a majority of students ($\geq 60\%$) demonstrating measurable improvement across competency constructs. Within 12 months of publication, the project will document adoption or substantive adaptation of the curriculum module by at least two additional ALA-accredited LIS programs beyond the pilot sites. The on-demand course will enroll at least 45 practitioners in its first year of availability, with a completion rate of 60% or higher. Open-access research deliverables will achieve at least 250 cumulative downloads within 12 months of publication. Dissemination goals include presentation at a minimum of two national professional venues and one regional venue.

PROJECT RESULTS

By the end of the project period, DECODE will have produced the first empirically grounded framework defining what LIS professionals must understand about **AI data** to make informed stewardship decisions. These results will equip LIS educators, graduate students, and practitioners with shared vocabulary, structured competencies, and tested instructional resources that enable responsible institutional decision-making about **AI data**.

Deliverables

The Statement of Need identified three interconnected gaps: the absence of frameworks for understanding what **AI data** are, inadequate knowledge among practitioners and students about **AI data** structures and risks, and the lack of curriculum resources for integrating this knowledge into LIS professional education. Each deliverable is designed to close one or more of these gaps in sequence. The Conceptual Framework establishes the foundational technical vocabulary the field currently lacks. The Competency Framework translates that vocabulary into practitioner learning needs. The curriculum module and on-demand course deliver that knowledge through both graduate education and continuing professional development. The journal article positions the profession to act on these findings by confronting the jurisdictional question that must be answered before any institutional commitment to AI stewardship can be made.

All deliverables are designed for broad adoption beyond the institutions directly involved in this project. Research deliverables will be deposited in IUScholarWorks and OER Commons under Creative Commons licenses, ensuring permanent open access. The curriculum module is structured for integration into existing LIS courses rather than as a standalone requirement, with competencies organized by practitioner role to support selective adoption across programs with varied missions and resource levels. The Competency Framework is explicitly designed for use by LIS programs, employers, and professional organizations beyond those involved in the project. The modular, open-access design of these deliverables positions them for potential adaptation to alternative credentialing pathways, including micro-credentials, by programs and organizations best positioned to implement them within their own credentialing contexts. The on-demand course extends practitioner access nationwide without requiring institutional travel or professional development budgets.

Conceptual Framework for Understanding AI Data (Month 8). A document characterizing the types and risk profiles of AI data, including training data, model architectures, learned parameters, configuration files, documentation, and adversarial and security-relevant artifacts, in terms accessible to LIS practitioners, directly addressing the foundational gap in the field's understanding of what **AI data** are and how they differ from other digital objects subject to archival stewardship. The framework will address preservation risk categories, including artifacts that may pose safety, privacy, or security concerns if retained or made openly available. Published under a Creative Commons license, it will serve educators, institutions, and professional organizations as a foundational reference.

Competency Framework for AI Data Stewardship (Month 14). An empirically derived framework organized around the knowledge, skills, and dispositions LIS practitioners need to engage meaningfully with **AI data**, structured to enable direct translation into curriculum learning objectives, employer position descriptions, and professional organization continuing education priorities. The framework will distinguish competencies appropriate to different practitioner roles (archivist, digital preservation specialist, generalist librarian) to support targeted adoption across varied program and institutional contexts.

Graduate Curriculum Module (Month 22). A tested, openly licensed module with learning objectives, readings, activities, and assessments organized around the three Competency Framework constructs, designed for integration into existing LIS courses rather than as a standalone requirement. Includes an instructor guide with implementation recommendations for varied program contexts. Deposited in OER Commons for open access.

Journal Article on LIS and AI Stewardship (submitted Month 24). A peer-reviewed manuscript synthesizing empirical findings from all three research questions to address the central question the profession has not yet confronted: whether LIS has the capacity and responsibility to claim jurisdiction over **AI data** stewardship, and what foundational knowledge would be required to do so. Submitted to the *Journal of Education for Library and Information Science* (JELIS) by Month 24.

On-Demand Course (Month 30). A self-paced professional development course on NEDCC's Training Hub, adapted from the pilot-tested graduate curriculum module and including a summative learning assessment with immediate learner feedback. Course format and scope will be determined collaboratively with NEDCC based on pilot results.

Practitioner Briefs (Months 9, 15, and 30). Three practitioner briefs translating project findings into accessible guidance for working professionals, submitted for publication in *Archival Outlook* and *American Libraries* and made permanently available on the project website. Brief content is described in the Dissemination section below.

Collectively, these results will (1) provide the profession with a shared conceptual vocabulary for understanding **AI data**; (2) reduce documented knowledge gaps among graduate students and practitioners; (3) provide validated instructional resources ready for national adoption; and (4) establish an evidence base for ongoing professional debate about AI stewardship jurisdiction.

Dissemination

LIS educators (ArchiPrEd SIG, ALISE conference, and AERI): A dedicated ArchiPrEd SIG session during ALISE 2027 (November 2027), coordinated through Dr. Buchanan as SIG Convener, will present Phase 1 findings and the Conceptual Framework, featuring the PD, Dr. Tao Dong, and Thomas Padilla, and gather structured feedback from the 75-member SIG community to inform Phase 3 curriculum development. ArchiPrEd SIG engagement will also occur mid-year through faculty survey pilot testing and distribution. A juried panel submission to ALISE 2028 (November 2028) will present complete project findings and provide implementation guidance to support curriculum adoption and adaptation across ALA-accredited programs, with a companion webinar or pre-conference workshop specifically targeting faculty implementation of the curriculum module. Findings will be submitted for presentation at AERI in Year 2, with potential follow-up engagement in Year 3 as curriculum results emerge. The journal article manuscript will be submitted to JELIS by Month 24.

Current practitioners (NEDCC, SAA, and regional venues): Dissemination to practitioners provides immediately usable guidance for appraisal and preservation decisions in active institutional contexts. NEDCC's Digital Directions conference will host a live presentation in Year 3 with 28 registration scholarships, followed by launch of the on-demand Training Hub course with 45 enrollment scholarships, awarded through an application process administered by NEDCC. The SAA Annual Meeting will be targeted in Year 2 to present Phase 2 findings on practitioner knowledge gaps to the broadest national constituency of archival professionals. The Midwest Archives Conference (MAC), which draws regional archival practitioners from across the Midwest and is accessible without extended travel, will be targeted in Year 1 for an early-stage presentation of the project's rationale and preliminary Phase 1 findings. The Society of Indiana Archivists (SIA) annual conference will provide an additional regional venue for dissemination to Indiana-based practitioners. Additional venues such as the Association of Tribal Archives, Libraries, and Museums (ATALM) conference may also be explored, where AI has recently emerged as a program subtheme. Three practitioner briefs will translate findings into accessible guidance for working professionals: Brief #1 (Month 9) will present the Conceptual Framework's key distinctions about **AI data** types and their stewardship implications; Brief #2 (Month 15) will report on the knowledge gaps identified in Phase 2 and their implications for practitioner preparation; Brief #3 (Month 30) will summarize curriculum findings and introduce the on-demand course. All three briefs will be submitted for publication in practitioner outlets including *Archival Outlook* and *American Libraries*. A free webinar (Month 10) will present the Conceptual Framework to the broadest possible practitioner audience, with recording made permanently available on the project website.

Graduate students: The PD will integrate project findings directly into IUB's MLS and archives and records management courses as they emerge, providing immediate pedagogical impact at the home institution and demonstrating a replicable model for adoption by other LIS programs. A project website hosted at IUB (launching Month 3) will provide open access to all deliverables, practitioner briefs, and recorded sessions throughout and beyond the project period.

Sustainability and Adaptability

Sustainability is built into the project design at four levels.

Practitioner education infrastructure: The on-demand course on NEDCC's Training Hub is the project's primary sustainability mechanism. Dedicating the third year to course development ensures the product benefits from the complete research and pilot testing cycle. IMLS investment funds the development and initial access costs; after grant-funded scholarships are exhausted, NEDCC will continue offering the course through its standard cost-recovery model, consistent with how NEDCC sustains all Training Hub courses beyond initial grant funding. This arrangement is not incidental to the project's design: it is how IMLS-funded practitioner education achieves sustainability without requiring continued public subsidy. Because **AI data** practices and documentation standards continue to evolve, the maintenance plan developed in Phase 4 establishes a structured review schedule to ensure the curriculum and course remain current over time. The openly licensed research deliverables ensure that future researchers and educators can build on and revise foundational materials as the field develops; when the course requires more than incremental updating, that need will constitute the basis for a future grant application.

Open access preservation: All research deliverables will be deposited in IUScholarWorks and OER Commons under open licenses. The on-demand course is a separately produced professional development product adapted from these openly licensed materials, which NEDCC will sustain through its standard Training Hub catalog.

Professional infrastructure: Periodic faculty review through ArchiPrEd SIG networks will support ongoing refinement of the curriculum module. **Curriculum design infrastructure:** The curriculum module is designed for integration into existing courses rather than as a standalone requirement, lowering barriers to adoption across programs with varied resource levels and curricular structures. Its openly licensed, modular format enables selective adoption of individual components, further reducing the institutional commitment required.

DECODE: Schedule of Completion — Year 1 (Months 1–12)

Developing Competencies for AI Data Evaluation in Library and Archival Education | Indiana University Bloomington | Year 1 of 3

Activity	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
IRB Application Submitted												
Phase 1: Systematic Literature Review												
Phase 1: AI Practitioner Interviews												
Phase 1: Document Analysis of AI Artifacts												
Project Website Launch												
Advisory Board Meeting #1												
Advisory Board Meeting #2												
Conceptual Framework [Deliverable]												
Phase 2: LIS Practitioner Interviews												
Phase 2: LIS Student Focus Groups												
Practitioner Brief #1 [Deliverable]												
Webinar: Conceptual Framework												
MAC Conference Presentation												

Deliverables are marked [Deliverable]. Advisory board meetings occur at Months 6, 12, 18, 24, 30, and 36.

DECODE: Schedule of Completion — Year 2 (Months 13–24)

Developing Competencies for AI Data Evaluation in Library and Archival Education | Indiana University Bloomington | Year 2 of 3

Activity	Month 13	Month 14	Month 15	Month 16	Month 17	Month 18	Month 19	Month 20	Month 21	Month 22	Month 23	Month 24
Phase 2: LIS Interviews (complete)												
Phase 2: Focus Groups (complete)												
Phase 2: LIS Faculty Survey												
Competency Framework [Deliverable]												
Formative Evaluation — Conway												
Practitioner Brief #2 [Deliverable]												
Phase 3: Curriculum Module Development												
Advisory Board Meeting #3												
SAA Annual Meeting Presentation												
AERI Conference Presentation												
Phase 3: Pilot Testing (IUB/UMD/Pitt/NCCU)												
Phase 3: Pilot Revision												
Curriculum Module [Deliverable]												
ALISE 2027 / ArchiPrEd SIG Session												
Journal Article Submitted [Deliverable]												
Advisory Board Meeting #4												

Deliverables are marked [Deliverable]. Advisory board meetings occur at Months 6, 12, 18, 24, 30, and 36.

DECODE: Schedule of Completion — Year 3 (Months 25–36)

Developing Competencies for AI Data Evaluation in Library and Archival Education | Indiana University Bloomington | Year 3 of 3

Activity	Month 25	Month 26	Month 27	Month 28	Month 29	Month 30	Month 31	Month 32	Month 33	Month 34	Month 35	Month 36
Phase 4: NEDCC Course Collaboration												
Phase 4: Course Production												
Advisory Board Meeting #5												
Practitioner Brief #3 [Deliverable]												
NEDCC On-Demand Course [Deliverable]												
Digital Directions Presentation												
Maintenance & Updating Plan												
Summative Evaluation												
ALISE 2028 / ArchiPrEd SIG Session												
Advisory Board Meeting #6												