Cultivating Equitable and Sustainable Ecosystems for Legacy Research Data

Rutgers University's School of Communication and Information, in partnership with Indiana University's Luddy School of Informatics, Computing & Engineering, requests $192,137 to support a two-year applied research project to study existing ecosystems of legacy research data and better understand the role libraries can take in managing older research data across disciplines and organizations. This project addresses a critical national need for libraries and researchers to have a robust framework to determine what data to keep, why, how, and for how long (Borgman, 2019). The proposed project will **examine existing efforts and decision-making regarding the value, management, and use of legacy data, directly informing and impacting open science and equity and inclusion practices in knowledge production.** It aligns with **Goal 3** of the National Leadership Grants for Libraries Program ("Improve the ability of libraries and archives to provide broad access to and use of information and collections with emphasis on collaboration to avoid duplication and maximize reach.") and addresses **Objectives 3.1 and 3.2** as it will contribute to enhancing digital infrastructures and services and increasing access to specific information and resources.

Legacy or heritage data rescue efforts involve migrating data from their original formats and contexts into trustworthy digital environments (Lin et al., 2020). There is a noted lack of common solutions as collections are different, and successful curation often depends on community-specific metadata and expertise (Sabharwal, 2021). For many relevant datasets created before or during the transition to digital technologies in the 1990's, experts most familiar with the data have retired or are soon to retire. This tipping point lends urgency to the proposed study and suggests a timely opportunity to explore the present value and potential of legacy data for catalyzing new research across disciplines and sectors. The proposed study responds to this pertinent national need while aiming for rapid impact across professional communities through applied research and dissemination.

1. Project Justification

Older research data, or *legacy data*, broadly defined here as data collected in the past that is often inaccessible to potential users, is an extremely valuable resource for scientific inquiry as it leads to new discoveries and changes in scientific paradigms (Heidorn, 2008; Stahlman, 2020; Smith, Keesstra & Rose, 2015). Robust climate modeling, for example, depends on data that goes back hundreds of years, with some data, such as tree-ring chronologies, going back thousands of years (Jones et al., 2009; Duncombe, 2023). In astronomy, data collected within a limited time range makes it difficult to conduct rigorous modeling of complex astrophysical objects, such as black holes and neutron stars (Vitale, Biscoveanu, & Talbot, 2022). Ancient DNA from well-preserved fossils and their digitized counterparts have allowed us to reconstruct the evolutionary history of many species, including dinosaurs, horses, and humans (MacFadden, 2005; Warren, 2019). Our knowledge of the world, its past, present, and future, depends on our ability to preserve, maintain, and provide access to older research data.

As the United States federal government has declared 2023 to be the Year of Open Science, improved preservation and access to legacy data becomes an important component of national open science policy and the efforts to accelerate discovery and innovation, promote public trust, and drive more equitable outcomes (OSTP, 2023). Legacy data is abundant, and as more of the original data creators retire and transfer their research data legacies to institutions, academic libraries will need to decide what to do with such legacies and how prioritize resources, training, collaborations, and access (Griffin, 2015). How can we (researchers, libraries, and society) better utilize the trove of data that has been collected in the past? Which efforts take priority and who needs to be involved in them? How can libraries help cultivate equitable and sustainable systems for legacy research data regardless of their media and institutional status of the creators? Answering these questions and creating wide-scale legacy data

migration and curation frameworks are imperative, and libraries are well positioned to do that by virtue of their stewardship mission and archival expertise.

Many data rescue initiatives were launched in recent decades, indicating the urgency, importance, and potential impact of preserving historical data in studying scientific phenomena over time. Table 1 below highlights several collaborative data rescue projects and their challenges and outcomes:

Table 1. Examples of documented legacy data rescue projects.

Discipline and Source	***Description***	***Challenges***	***Outcomes***
Social sciences (Altman, 2009)	Years of analog data from longitudinal studies, particularly of women and vulnerable populations	Organizational transformation and merger, aging media, obsolete and cumbersome technologies	Improved archival and discovery, best practices in digital preservation, collection management strategy
Astronomy (Grindlay et al., 2011)	Large-scale project to digitize a century of astronomical plates	Slow hardware and software, transcription of scanned logs, processing of special requests	Ongoing digitization of the collection, publications using data from the archive, citizen science engagement
Meteorology / Climate science (Mayernik, et al., 2017)	Data from historical print weather records	Diverse geographic and temporal dimensions of data, sparse metadata, obsolete formats, and technologies	Improved and standardized metadata, assessment, and processing workflows
Ecology (Specht, et al., 2018)	Vegetation survey plot data beginning in the 1880s from published texts and other sources	Heterogeneous and disaggregated data, aging media, lack of standardized formats	Improved and standardized metadata, data digitization and integration, documented socio-technical challenges
Geoscience (Wippich, 2012)	Images, documents, and historical records on geography and earth sciences	Diversity of data, obsolete formats and technologies, inconsistent metadata, storage	Digitized data and metadata, improved access to legacy geoscience data

As can be seen from the summary table above, legacy data initiatives range from prioritizing digitization of analog data (astronomy, ecology, geoscience) to the development of strategies and techniques for the future (social sciences, climate science). There are many valuable lessons that can be learned from each of these initiatives, but their diversity makes it difficult to draw systematic conclusions about the value of legacy data and what working with and preserving legacy data means to various communities. A systematic study that synthesizes the lessons and insights from these and other separate initiatives and creates a map of the legacy research data ecosystem can be an

effective tool to guide future efforts. The applied research project proposed here aims to fill this gap and to **a) examine the perceived value of legacy data and the division of roles and responsibilities in legacy data management, and b) develop a systematic understanding of legacy research data efforts across disciplines and stakeholder communities.** The project will engage with discourses and communities that are involved in legacy data preservation and management and examine the socio-technical processes and impact of working with and curating legacy data, as well as the efforts to promote equity, fairness, and justice in and through legacy data.

The project builds upon the data rescue efforts described above and discipline-oriented efforts to develop rescue data assessment frameworks (Downs and Chen, 2017; Mayernik et al, 2017; Shiue et al. 2021). It will enhance the collective understanding of legacy data value by examining the vast literature on legacy data and identifying and comparing practices and attitudes across two case studies. As a systematic work that synthesizes multiple discipline-oriented insights in a well-scoped systematic investigation, the project differs from prior work in three important ways: 1) It focuses on understanding how various stakeholders, especially libraries and researchers, perceive the value of legacy data; 2) It works with two diverse case studies that have deep concerns over equity and inclusion and will enable cross-disciplinary comparisons; and 3) It prioritizes interdisciplinarity and collaboration and incorporates both research and educational outcomes as equally important aspects of legacy research data efforts.

In addition to addressing gaps in our knowledge about the value and practices in legacy data curation, the proposed project also addresses two other issues crucial for research that aims to be relevant for library and archival practice: **the need for curricular and training materials** for librarians and data professionals and **the need for partnerships and collaboration** in libraries that is crucial if libraries are to remain visible and relevant (Heidorn, 2011; Lyon, Patel & Takeda, 2014; Mayernik et al., 2017). The project will develop robust curricular materials to support legacy data management and publish them in open repositories for educators and librarians, therefore contributing to the growing field of digital data curation and its skills and competency requirements (Higgins, 2018; Poole, 2016; Palmer et al., 2013; Kim, 2015). Through the case studies and dissemination in data management communities, this project will promote awareness of collaborative models that enable libraries to build cross-university partnerships, increase buy-in for all involved, and establish an infrastructure for ongoing and future projects (Latham and Poe, 2012).

The increasing momentum of open science and widespread adoption of the principles of findability, accessibility, interoperability, and reusability (FAIR) and the principles of collective benefit, authority to control, responsibility, and ethics (CARE) to support data reuse make this project a timely opportunity to explore the present value and future potential of legacy data for catalyzing new research across disciplines and for improving libraries' services and collaborations.

2. Project Work Plan

Key challenges in managing legacy data include inadequate infrastructures and support as well as lack of specialized equipment needed for data migration. Furthermore, legacy data may represent exploitative Western science practices from the past, such as data collection at the expense of local communities. These challenges motivated the design of this project and the development of research questions, themes, and activities.

The project addresses the following research questions:

- RQ1) How is the value of legacy data perceived and constructed across research and professional communities?
- RQ2) Who is involved in legacy data preservation efforts?

- RQ3) What forms of collaboration work best in curating and preserving legacy data?

These questions will be interwoven with three integral themes that will guide the study's theoretical and methodological approach, allowing for more nuanced exploration of the research questions:

Theme 1: ***Legacy data for open science*****.** This theme highlights the importance of linking historical data to journal publications, research proposals, and other resources through evolving open science technologies and cyberinfrastructures (Gentemann, Erdmann & Kroeger, 2022). It will help identify existing patterns in how various communities perceive and handle older data [RQ1], including the institutional structures and collaborations [RQ2], and opportunities for creating and maintaining linkages across communities [RQ3].

Theme 2: ***Sharing equipment and knowledge*****.** This theme addresses the challenges of working with specialized and often antiquated equipment that is needed in research data migration (Pevtsov, et al., 2019), and the degrees to which institutions and individual researchers are willing to share their valuable physical and intellectual resources [RQ1, RQ2]. It will guide the development of research instruments in emphasizing collaboration and sharing, with a particular emphasis on inequities in access and use [RQ3].

Theme 3: ***Engaging local communities*****.** Many disciplines have a history of colonialism while collecting data in locations of geographic interest (Lehuedé, 2021; 2022). For example, indigenous and local communities have argued that astronomers exploit the lands upon which they place telescopes and collect data [RQ1]. Moreover, under-resourced researchers may not have sufficient access to digitized data (Stahlman, 2023) as such access requires cyberinfrastructure they do not have [RQ2, RQ3]. Our research will pay particular attention to these aspects of legacy research data ecosystems and examine how to best diversify management, access, and use of legacy data [RQ3].

We will address these themes and questions by implementing a mixed-method approach (see Fig. 1).

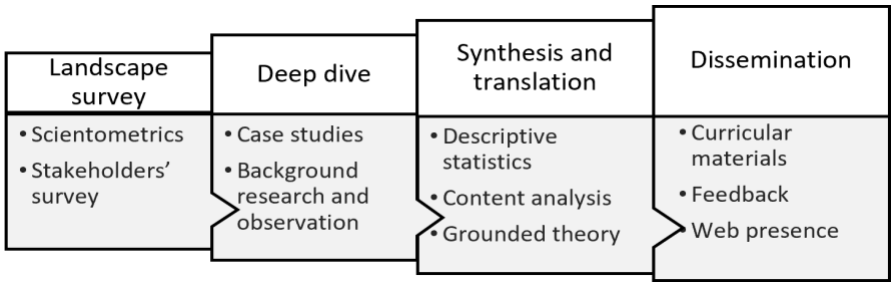

Figure 1. Mixed-method study design.

First, we will survey the landscape of legacy data efforts that will include an analysis of a corpus of representative publications and projects and a survey of librarians, archivists, researchers, and data managers aimed at understanding current attitudes toward and activities around research legacy data. Second, we will engage with two legacy data migration and preservation case study sites in astronomy and political science for focused investigation into underlying processes, technologies, and broader socio-technical and communication ecosystems. The findings from both the landscape survey and the deep dive will be analyzed using the methods of statistics, content analysis, and grounded theory and synthesized into a conceptual framework and shared with professional communities for peer feedback. Finally, our activities and findings will be translated into open educational resources that we will develop, pilot in LIS courses, and share with the data management networks for further dissemination and use.

I: Survey the Landscape of Legacy Research Data and Perceptions of the Wider Community.

Corpus analysis. Using online databases such as Google Scholar, Web of Science, and Scopus and the PRISMA guidelines for screening, selecting, and reporting search results (Page, et al., 2021), we will conduct a literature analysis of publications on the topics of legacy research data use, preservation, curation, and management. To

address Theme 3 of our project ("Engaging local communities") and the international context of our case studies described below, we will include non-English language publications in our search, focusing on the literature from the Global South and its underrepresentation in the Western indexing databases (Khanna, Ball, Alperin & Willinksy, 2023). We will annotate and qualitatively analyze these papers, looking for references to individuals involved, processes, data types, formats, and locations, and problems being addressed, as well as information relevant to our three guiding themes. Using quantitative scientometric and content analysis methods, we will also create a network of authors, data sources, and institutions involved in legacy data activities globally and identify temporal and terminological patterns across those networks.

Stakeholders' survey. Informed by the results of the literature analysis, we will create an online questionnaire for broader reach and wider insight into our research questions. The survey will be designed to examine perceptions of the value of legacy data, current and ideal roles taken in managing legacy data, and current and ideal forms of collaboration. We will include survey questions aimed at understanding underlying organizational, institutional, and disciplinary features and barriers in access to legacy data. To successfully reach our population of interest (librarians and archivists, researchers, and data management communities) we will leverage listservs and groups such as the Association of Research Libraries (ARL) for librarians, the Research Data Alliance for data management, and research organizations relevant to our case studies such as the American Astronomical Society (AAS) and the International Association for the Study of the Commons (IASC). Survey data will be analyzed using statistical and qualitative methods.

II: Deepen Landscape Understanding through Case Studies.

Case studies. We have identified two cases that provide rich context for our study. They represent active involvement of librarians, researchers, and data managers in legacy data efforts, commitment to open science, serious challenges for curation and preservation, and tensions (implicit or explicit) between local and research communities.

The Legacy of Elinor and Vincent Ostrom / The Commons Research: This case centers around physical and electronic artifacts of Elinor and Vincent Ostrom and their life's work as it is related to political science, economics, and the commons research. Elinor "Lin" Ostrom (now deceased) was one of the 2009 Nobel Prize Laureates in Economics for her research into how people can create rules and institutions that allow for the sustainable and equitable management of shared resources ("Elinor Ostrom: Facts", 2009). In 1973 she and her husband Vincent Ostrom founded The Ostrom Workshop (originally as the Center in Political Theory and Policy Analysis). Both researchers were exceptionally influential and prolific. Over the decades, they generated vast amounts of data through their collaborations across multiple disciplines, and left artifacts that include analog and electronic records and research data that are stored in boxes and databases. Some of those resources have already been lost. The challenges of this case include difficulties separating data from the rest of the legacy, lack of consistent descriptions of the assets, the scattered nature of relevant information that can help with research data organization and access, and loss of library ownership of some of the data assets. Systematizing the legacy of the Ostroms and their collaborations and learning about their efforts to provide wider and equitable access to data will contribute to the advancement of open science and sustainable data curation across the social sciences.

The National Optical-Infrared Astronomy Research Laboratory (NOIRLab) / Astronomy: This case focuses on telescope data in FITS (Flexible Image Transport System) file format currently stored on exabyte tapes at the NOIRLab in Tucson, Arizona. Each tape represents a night of observations at an observatory telescope in United States or Chile conducted between 1994 and 2004, and the collection may contain roughly 4 million

individual files. Prior to a recently launched curation effort (Steffen & Hunt, 2022), the data were not cataloged or accessible online, and the antiquated media reading equipment needed to migrate these data into a modern computing environment is becoming harder to acquire as time passes. The challenges include incomplete metadata, loss of cultural and institutional knowledge as scientists most capable of assisting with metadata are approaching or past retirement, and a potential loss of data due to degradation or slow progress. Making this collection widely available to researchers worldwide will contribute to larger efforts at NOIRLab and throughout the astronomy community to open and democratize data.

PIs Stahlman and Kouper have deep long-term connections with these case studies. Their personal and professional relationships with data managers and researchers at the Ostrom Workshop and the NOIRLab will help to establish trust with broader networks and gain access to participants who otherwise might be unwilling to provide input (especially, in the context of Theme 3 of our project). Letters of support from the leadership of each case study (McManus and Raymond) confirm their enthusiasm and willingness to participate in this research. For each case study we will visit their main locations in Arizona and Indiana and engage in observations and interviews to learn about the projects and their participants and data.

Background research and observations. In addition to immersing ourselves in these case study sites for focused subject and disciplinary insight, we will approach our cases from an interpretive and contextual inquiry perspective in order to understand and be attentive to what actions, discourses, and events mean to case study participants (Geertz, 1980). We will collect relevant documents, artifacts, and notes that will allow us to form a multilayered description and document assumptions, expressions, and meanings of these organizations and their data work. While true ethnography is not practical within the timeframe of this project, we will remain in close contact with site representatives and their legacy data activities as we collect and interpret qualitative findings.

III: Synthesize Findings and Translate Research Findings into Practical Outcomes.

Through Phases I and II this project will implement a mixed-methods strategy to collect diverse data through document search, surveys, interviews, and observations. In Phase III we will synthesize data and findings of our study by using statistical and content analysis methods. We will generate descriptive statistics that illustrate stable patterns and trends in legacy data curation and identify major stakeholders and their contributions. We will use a grounded theory approach (coding-memoing-sorting-writing) and triangulation to develop an in-depth, trustworthy insight into our research questions and themes, increase the credibility of our findings, and generate a unified representation of data, methods, investigators, and environments (Denzin, 2012; Heale and Forbes, 2013).

To create a conceptual framework that will allow us to synthesize the findings and translate them into more practical outcomes (see Project Results below), we will map the concepts and processes identified in Phases I and II and iterate over their categorization and re-categorization in order to arrive at a valid framework that fits with our data and constraints (Jabareen, 2009). The findings will be then discussed with our case study participants and major stakeholders to further validate and collectively confirm our findings, the framework, and ethical orientations through a dialogic approach (Harvey, 2015).

Our activities and findings will be translated into Open Educational Resources (OER) that we will develop and pilot in our own courses (such as Digital Curation at Rutgers U and Management, Access, and Use of Big and Complex Data at Indiana U). Curricular materials will fall into three categories. First, we will compile our insights into an annotated bibliography about legacy data with a theoretical introduction. Second, we will construct short hands-on activities surrounding the data lifecycle, to be published as online tutorials and modules that can be incorporated easily into classrooms and library training programs. Third, we will develop and publish longer-term activities for

use as course projects. Educational and training materials will be published in OER Commons and as Library Carpentry lessons. Curricular materials will also be promoted at conferences and through LIS channels such as the JESSE listserv as part of our dissemination plan (Phase IV).

IV: Disseminate Knowledge and Establish Feedback and Improvement Loops.

To ensure that we address all three themes of our project appropriately and in depth and incorporate them into our instruments and analyses, we have convened an expert advisory board that includes experts in digital preservation, international science and innovation, and research data management (see letters of support):

- **Rebecca Koskela**, Executive Director of Research Data Alliance US, who brings substantial expertise in research data management and cyberinfrastructure (Stall, et al., 2018). Koskela's guidance will be especially relevant to Theme 1 ("Legacy data for open science").
- **Sarah Buchanan**, Associate Professor in Library & Information Science at University of Missouri, who has engaged extensively with data and cultural heritage rescue efforts and LIS education (Buchanan, 2022). Buchanan's guidance will be especially relevant to Theme 2 ("Sharing equipment and knowledge").
- **José Guridi**, Ph.D. student in Information Science at Cornell University, with prior professional experience developing data infrastructures and science policy in Chile (Guridi, Pertuze, & Pfotenhauer, 2020). Guridi's experiences and expertise will be invaluable in addressing Theme 3 ("Engaging local communities").

Through meetings that will be held at least once a year with more check-in meetings added as needed (at the beginning of the year and toward the end of the year), the board will provide guidance in every aspect of our project, including data collection and analysis, development of curricular materials, and dissemination. The Advisory Board will also be instrumental in expanding our networks and venues for knowledge dissemination.

In addition to dissemination of educational and training materials through specialized venues, project outcomes will be disseminated through conference presentations and publications targeting three audiences: information science researchers, data curation practitioners (librarians, archivists, and data managers), and other legacy data stakeholders (researchers). To reach the information science researchers audience, we report on research findings via peer reviewed journal articles and through presentations at conferences such as the Association for Information Science & Technology (ASIS&T) and iConference. Where possible and in consideration of human subjects research protocols, project data will be disseminated for reuse through platforms such as the Open Science Framework (OSF) and institutional repositories, per our data management and digital products plans.

To reach the data curation practitioners audience, we will attend curation and archival conferences, including the Research Data Alliance (RDA) Plenary Meetings, the International Digital Curation Conference (IDCC), and the Society of American Archivists annual meetings. We will also collaborate with the Data Curation Network (DCN, see letter of support) to disseminate our findings and solicit peer feedback.

We will also attend disciplinary conferences related to our case studies to disseminate project findings to researchers who are involved in legacy data work, such as American Astronomical Society (AAS), Library & Information Services in Astronomy (LISA), and The Workshop on the Workshop (WOW) Conference at Indiana University. This approach will allow us to interface with researchers and data managers working within particular research communities and, again, broaden our dissemination networks.

To increase the project's visibility and solicit broader feedback from relevant communities, we will implement and maintain a web presence for the project and its participants. With the permission of our case studies, each case will be featured on the website. The project website will record the outcomes of the project beyond its grant-funded

timeframe leading to additional engagement and community building. The online web presence will be supplemented by team members' social media activities that will be geared toward feedback and community engagement.

Progress and success of the project will be tracked through regular meetings and monitoring of our key results and deliverables (see Performance Measurement Plan and Schedule of Completion for further details). Specific coordination efforts will include regular meetings to plan data collection activities, analyze results, and write manuscripts. Our stepwise research design and detailed timeline will help to ensure that each research activity can be assessed independently and complementary to other activities. Project success will be measured through the number and reach of publications, presentations, and documentation; the number of students and collaborators engaged in the project, and project-level metrics specific to each case study site. The project will be managed and executed by the Project Director Stahlman and co-Director Kouper, who will collaboratively coordinate all activities, including data collection and analysis, case studies coordination, curricular and community development, and student mentorship and training.

3. Diversity Plan

The proposed project is relatively small in scope and size, limiting potential impact on critical diversity issues. Nevertheless, it is designed, led, and will be implemented by an all-women research team who have a record of paying attention to gender and other diversity issues (Herring et al, 2006; Kouper, 2010; Kouper, 2011). Co-PI Kouper has also been an active member of several national and local organizations that advance gender equity and contributes to the undergraduate and graduate mentorship programs that encourage first-generation and underrepresented students to participate in research. Our case studies were also carefully selected to highlight and deconstruct key challenges for research communities such as gender imbalances in science and marginalized access to valuable research resources such as data.

Through its Theme 3 the project will examine connections between local and indigenous knowledge and legacy data and approach relevant use cases with utmost respect for cultural protocols. The project will also empower researchers and knowledge institutions to preserve potentially valuable historical research data and draw attention to how local communities and marginalized researchers can engage with these data. A multilingual graduate student assistant, preferably from Latin / Hispanic background, will be recruited to assist with the project, allowing us to adopt a global focus and de-center English language literature and documents through the analysis process. All collaboration and dissemination efforts will prioritize under-resourced and underrepresented institutions and communities.

Through case study site immersion and advisory board engagement (with two female and one Latin-American membership), a feedback loop will be created to ensure that the diverse voices of stakeholder-participants are integral to the evolution of the project from beginning to end and beyond. Furthermore, the curricular materials that we create will be openly available and promoted through targeted outreach for the benefit of less resourced programs and minority-serving institutions. We will pay special attention to the issues of inclusiveness and underrepresentation when working with graduate student assistants and collaborators and soliciting diverse opinions for feedback. Members of the advisory board are uniquely positioned to advise us on this through their own experience and expertise.

Finally, ageism is a known issue and barrier to DEI. Legacy data typically originates with previous generations of data producers, who may remain unsupported in their goals for making data more widely available, often as personal and professional legacy (Stahlman, 2022). A key objective of this project is to work closely with those

most familiar with the data where possible, and to engage older generations, especially female and foreign stakeholders, in curation.

The following principles will guide our choices and decisions:

- Incorporate DEI themes in our research and curricular activities
- Disseminate findings to smaller institutions that may have fewer resources, encourage their faculty and students to collaborate and engage in scholarly activities
- Engage diverse students in library science, informatics, data science, and similar domains and provide mentorship and learning opportunities through project activities
- Create an inclusive and intellectually diverse working environment
- Consider alternative forms of assessment, collaboration, and dissemination, including altmetrics, social media, virtual environments, and open platforms

As white persons, we recognize that we benefit from the privileges and opportunities that come with being part of the dominant group in society. Through our own concerted effort and studies, we have learned about the impact of systemic discrimination and have come to understand the importance of actively working towards greater diversity, inclusion, and equity. In addition to the proposed activities that are part of applied research in this project, we will seek out opportunities to engage with diversity and inclusion by participating in workshops and training at our universities and by working to create inclusive spaces that go beyond this project.

4. Project Results

The intellectual outcomes of this proposal include: 1) Value assessment of legacy data and researchers' needs in several disciplines as well as impacts on local communities globally; 2) Understanding of the current approaches to curating and preserving legacy data in academic libraries in the US; and 3) Open curricular materials for training LIS and data professionals. Project outcomes will augment the collaborative capabilities of academic libraries, reduce barriers to legacy data access and curation, and inform strategic partnerships for the development of cyber- and social infrastructures.

The primary knowledge gap to be addressed by this project is the fundamental value of legacy research data and decision-making guidance. Saving all data produced over time is impractical, but the loss of important resources can have unanticipated effects and consequences. As an economic concept, "value" refers to benefits of goods and services, while more broadly the term can refer to practical utility, and even to individual and societal values, esteem, and legacy itself. Therefore, assessing the value of legacy data (as defined in this proposal) is a complex endeavor, especially considering that the designation "legacy data" encompasses a broad range of data types, formats, locations, and contexts. Through quantitative and qualitative analysis and collaborative feedback, the research proposed here will identify nuanced measures and indicators of value and engage stakeholder communities in developing informed strategies that can be applied across communities for assessing and enhancing the value of data. The breadth of intellectual outcomes produced by this project is presented in Table 2 below.

Table 2. Intellectual outcomes by research question.

Research Question	Outcomes

RQ1) How is the value of legacy data perceived and constructed across research and professional communities?	• Aggregate perceptions, measurements, assessments of value and methods of its enhancing • Organizational, institutional, and disciplinary drivers in the international context • Community-specific perceptions and knowledge • Understanding of the role of infrastructure (equipment and instrumentation) and research data production and its inequities in access and use
RQ2) Who is involved in legacy data preservation efforts?	• Review of the current landscape and common roles taken by each population; effectiveness of these roles in relation to data curation and impact • Opportunities and challenges for existing and prospective roles • Insight into and critique of the existing inequities in access to knowledge and data in legacy data ecosystem (esp. in the context of Global South) • Synthesis of training and curricular materials that incorporate the themes of open science, equipment, and local communities
RQ3) What forms of collaboration work best in curating and preserving legacy data?	• Understanding of current forms of collaboration • Methods of assessing forms of collaboration in relation to organizational, institutional, and cultural structures and norms • Strategies for broader communities and contexts across disciplines

The second gap in knowledge to be addressed by this project is appropriate methods for training in issues, decisions, and techniques specific to legacy data curation and management, as well as a lack of portable curricular modules and syllabi that can be used for digital curation education across LIS and data management programs. To this end, the iterative strategy of applied research and community engagement is critical. Other information fields such as knowledge organization and data science have strong involvement of research communities in developing shared pedagogical strategies and teaching materials (Glushko, 2013; Oh, et al., 2019). For digital curation, such a community is still emerging, and areas of focus and even terminologies vary (Costello, 2010). Our curricular activities will cultivate shared digital curation education practices and community through the hands-on learning models and instructional modules that we will develop and test through our legacy data case studies.

Finally, the third knowledge gap we will address is how academic and research libraries can best engage and collaborate with research communities to fully operationalize the value of legacy data as well as their own capacities for leadership in this area. Building upon our assessment of the perceptions of value of legacy data across stakeholder groups (Phase I) and subsequent case study immersion and synthesis (Phases II and III), we will acquire extensive understanding of the challenges and affordances of current forms of collaboration. We will use this understanding to produce strategies that can be leveraged across disciplines and contexts (Phase IV).

The outcomes of this research including research findings, protocols and policy recommendations, and educational materials will cross-benefit academic and research libraries, research communities and scientists, and LIS educators and students.

Applicant Name: Dr. Gretchen Stahlman, Rutgers University
Project Title: Cultivating Equitable and Sustainable Ecosystems for Legacy Research Data

Schedule of Completion

Timeline of Activities and Outcomes	Year 1					Year 2			
	Q1	Q2	Q3	Q4		Q1	Q2	Q3	Q4
Pre-Phase: Planning and compliance									
PI team virtual kick-off and planning meeting	X								
Obtain IRB approval	X								
Phase I: Survey the Landscape									
Bibliometric analysis of scholarly literature:									
Sample and organize data in shared environment	X								
Establish coding scheme and train student	X								
Coding of papers (content analysis)	X	X	X						
Citation, authorship, and topical networks			X	X					
Survey stakeholders:									
Sample stakeholder populations	X								
Develop questionnaire	X								
Pilot questionnaire	X	X							
Distribute questionnaire			X						
Analysis of survey data (qualitative, statistical)			X	X		X			
Phase II: Deepen Understanding									
Case studies set up			X						
Case study research:									
Explore sites – background research				X					
Observations, interviews, and analysis				X		X			
Phase III: Synthesize Findings									
Data synthesis and triangulation				X		X	X	X	
Curricular development:									
Construct learning material based on Phase I-II				X		X	X	X	
Pilot curricular material at Rutgers and IU							X	X	
Phase IV: Dissemination and Community Feedback									
Advisory Board virtual meetings	X					X			
Finalize results and outcomes:								X	X
Establish/maintain web presence	X	X	X	X		X	X	X	X
Release open curricular material								X	X
Present at conferences		X		X			X		X
Publish scholarly research papers		X		X				X	X
Finalize sharing digital products and data	X	X	X	X		X	X	X	X

Applicant Name: Dr. Gretchen Stahlman, Rutgers University
Project Title: Cultivating Equitable and Sustainable Ecosystems for Legacy Research Data

Digital Products Plan

SECTION I: DIGITAL CONTENT, RESOURCES, OR ASSETS

In the course of this project the following digital content will be created:

1. Project website.
2. Online manuals and education and training materials.
3. Publications (journal articles, conference presentations, online materials such as social media and blog posts)
4. Code to process and analyze data (for bibliometric analysis and survey results)
5. Research data products (see Data Management Plan for more details)

This project does not focus on creating digital content beyond project online presence, research studies data and documentation, and training and community feedback materials (presentations and reports). We will use computers and software at Rutgers and Indiana University. The following file formats and software will be used:

- Website: HTML, CSS
- Online training material and documentation: MPEG, PDF, MS Office suite
- Publications: DOC, HTML, PDF, MS Office suite
- Code: Python, R, also NVivo for content analysis

To document digital products, we will use a simplified metadata approach. We will describe the resources in a readme file, documenting authorships, dates of creation and update, and other descriptive metadata. Digital products that will be preserved through Rutgers and Indiana University institutional repositories (preprints and data) will be described according to the standards of those repositories. Metadata will be managed as part of the digital asset maintenance and considered part of the digital assets and research objects and migrated together as needed.

Documentation and implementation processes and products will be regularly evaluated for completeness and consistency. The team will meet periodically to discuss quality issues and ensure that the work and products are aligned with the proposed scope. An additional quality check will be performed regularly during meetings with research participants and community stakeholders.

Project digital assets will be maintained using the Open Science Framework (OSF) or another similar platform as well as Rutgers and Indiana University storage systems (OneDrive and Microsoft Teams and Google Drive). Products of research (publications, data, presentations) will also be shared through traditional publication venues with priority given to open access venues. As the project is not directly creating digital collections and platforms beyond the project website and training materials, we will rely on the existing channels for widespread dissemination and use, including partner organizations and their websites, open publication venues, conferences,

listservs, and institutional repositories. Digital products will be openly available online, unless restricted by the publishing entities or limited by the requests of confidentiality.

SECTION II: AVAILABILITY AND ACCESS

The formal products of this project, including research studies data and documentation, training materials, and dissemination materials (website, presentations, publications) will be made available online unless restricted by the requests of privacy and confidentiality. We also anticipate intermediate products emerging as a result of conducting our work and documenting our progress. The formal materials and software products resulting from this effort will be licensed using open and free licensing, e.g., Creative Commons and Apache 2.0-style licenses. Intermediate products will be discarded by the end of the project life or re-used to become part of the formal products. We will adhere to FAIR guidelines in maximizing access to our formal products.

Working closely with case study sites, we will produce digital products related to those sites, including specific documentation, websites (or their updates), legacy data digitization, and so on. We will respect data management protocols of the case study sites in access and availability of digital products, while collaborating with them to suggest improvements as needed and incorporate best practices, including using open and sustainable formats and platforms.

Our team will also work with Rutgers and Indiana University data stewards to ensure that products developed in this project can be shared openly without violating any confidentiality or sensitivity policies (for example, if the workflows enable storage and transfer of restricted or embargoed data). The associated training materials will be shared broadly via open software repositories (such as OER Commons) and Rutgers and IU sharing platforms.

SECTION III: SUSTAINABILITY

Aligned with the fundamental objective of this proposal to cultivate a deeper understanding of long-term information management, we will adopt a “purposeful work” stewardship approach to our own data and digital projects (Palmer et al, 2013). For economic sustainability and risk management, we will rely on trustworthy, community-supported, and open platforms such as the institutional repositories managed by Rutgers and IU. To the extent possible, we will uphold the FAIR principles for all digital products and utilize file formats and metadata standards that are conducive to digital preservation. We will thoroughly document our processes and products for transparency. To further ensure sustainability of our digital products beyond the grant-funded period, our plan for succession is two-pronged. First, we will publish research products according to scholarly communication best practices. Second, we will ensure that informal products such as curricular materials are published in appropriate repositories with DOIs.

DATA MANAGEMENT PLAN

"Cultivating Equitable and Sustainable Ecosystems for Legacy Research Data"

This project will gather the data through observation, interviews, and an online survey in the form of audio recordings, transcripts, notes, and survey responses. Additional data will be gathered for bibliometric study, including metadata of sampled publications, annotated journal articles, and text notes.

During the course of the project, the following datasets will be produced:

- Dataset 1: Audio recordings of interviews (formats such as .wav, .mp3, etc.)
- Dataset 2: Transcripts of interviews (format such as .txt or .docx)
- Dataset 3: Survey instrument and responses (formats such as .csv and .docx for the codebook)
- Dataset 4: Metadata describing sampled publications (formats such as .csv or .xlsx)
- Dataset 5: Annotated journal articles (formats such as .xml or .csv)
- Dataset 6: Participant observation notes (formats such as .txt or .docx)

The data will be analyzed using grounded theory approach, qualitative coding, and quantitative scientometric and content analysis. Data will be collected and analyzed iteratively throughout the duration of the project.

Final products of the project include (see Digital Products Plan for more details):

- Publications
- Posters and presentations at scientific conferences
- Teaching materials (e.g., presentations, webinars, syllabi, and exercises)
- Public communication materials (e.g., social media, blog entries, etc.)

Internal review: Data collection involves human subjects and requires IRB approval. IRB application will be prepared and submitted when / if the project is approved for funding.

Sensitive information: Access permission to the data will be controlled through project-level permissions by the Principal Investigators (PIs) through the Open Science Framework (OSF) or another open collaboration platform that provides rigorous security controls for setting permissions to access a particular project. The PIs will have authority on which collaborators have access to each dataset and when. Following publication, the PIs will use the OSF and will create DOIs for anonymized data to allow public access when possible. The data will be available to both PIs and the research team involved in this project at every stage by accessing a private project on OSF. During the peer-review process following completion of the research, editors and reviewers may have blinded access to datasets on the OSF as needed and appropriate. De-identified survey data will be published in the appropriate repository such ICPSR or institutionally supported Figshare. Identifying information related to research participants that will be privately stored includes participants names, email addresses, affiliations, and other information deemed personal. Coded qualitative data will be stored securely and not shared outside of the project unless it is properly anonymized.

Technical dependencies: Data will be shared in open text-based formats so that any software can be used to analyze it.

Documentation: Codebooks and data notes will be created as part of data analysis (e.g., in the thematic coding procedures and in processing the surveys). Codes, their descriptions and other documentation that describes when, where and how the surveys, interviews, and observations took place will be stored in text formats along with the data. The documentation will be associated with the datasets through consistent file naming and through identifiers that refer to each data collection effort separately.

Data management and preservation: Data will be kept in a platform such as the Open Science Framework (OSF) and backed up separately on a password-protected external hard drive for the duration of the project and beyond. The OSF is a cloud-based repository for researchers (free up to 50GB). Data stored on the OSF is backed by a preservation fund that will provide for persistence of data, even if the Center for Open Science runs out of funding. The code base for the OSF is entirely open source, which enables other groups to continue maintaining and expanding it if the Center for Open Science is not able to. Folders with appropriate permissions for data, IRB documentation, and publications can be created in the OSF for active work with the data. For dissemination, we will use the OSF or a similar platform, as well as institutional repositories at each institution (RUCore and IU Scholarworks), and appropriate repositories such as ICPSR for survey data. Data publication involving human subjects will be limited to fully anonymized data (e.g., surveys). For other qualitative data (interviews, observation, etc.), we will rely on consent and guidance from subjects to make final decisions about data publishing.

Repositories to be used: Open Science Framework; RUCore; IU Scholarworks; ICPSR

Data management plan implementation and review: Adherence checks to the study's data management plan will occur every six months by both PIs to ensure the safety of the data collected and the privacy of research participants within the framework of the informed consent.

Applicant Name: Dr. Gretchen Stahlman, Rutgers University
Project Title: Cultivating Equitable and Sustainable Ecosystems for Legacy Research Data

Organizational Profile

Library and Information Science Department (LIS)
Rutgers University's School of Communication and Information (SC&I)

Statement of Purpose

Rutgers, The State University of New Jersey, is a leading national research university and New Jersey's pre-eminent public institution of higher education. SC&I's research and teaching focus on library and information science, organizational communication, social and new media, journalism and media studies, health communication, and information technology. A member of the Big Ten Academic Alliance, SC&I's programs prepare students for challenging careers in today's digital environment. The Middle States Commission on Higher Education voted June 21, 2018, to reaffirm the accreditation of all Rutgers University locations for July 1, 2018 to June 30, 2027.

Governance Structure

The Library and Information Science Department is located within the School of Communication and Information building, which is adjacent to the university's Archibald S. Alexander Library, the main social sciences and humanities library in New Brunswick, providing SC&I students and faculty with a bridge of knowledge to a 26-library system dedicated to education and information access. LIS is one of the three departments within the school whose 237 faculty instruct 2,186 full and part-time undergraduate majors and minors; 872 full and part-time master's students; and 50 doctoral students. Operational oversight is provided by the department chair, SC&I dean, and Office of the Chancellor-Provost.

Service Area

As a land grant university, Rutgers has a presence in all 21 of New Jersey's counties and annually enrolls almost 70,000 students collectively at Rutgers' three main campuses: New Brunswick (50,804), Newark (12,168), and Camden (6,569). Eighty-one percent of the students are NJ residents. Some key demographic indicators of the state include:

- Population: 9.2 million
- Median age: 40
- Median household income: $85,245
- Foreign-born; naturalized U.S. citizens: 31.2%
- Bachelor's degree or higher: 40.7%
- People with an internet subscription: 88.1%
- Ethnicity and Race: White alone: 55%; Hispanic or Latino: 21.6%; Black or African American alone: 13.1%; Asian alone: 10.2%; another race alone: 11.3%

History

LIS has a distinguished record of research, education, and community outreach that spans many decades, beginning with its origin in 1953 as the Graduate School of Library Service. In 1982, as the Graduate School of Library and Information Studies, it merged with the School of Communication Studies to form the School of Communication, Information and Library Studies (SCILS). The school now offers an Information Technology and Informatics major; a Master of Information with concentrations in Library and Information Science, School Librarianship, Data Science Informatics, and Design Technology, Information and Management, and Archives and Preservation; an ITI/MI Dual Degree Pathway; an interdisciplinary Master of Health Communication and Information; and a LIS-centric Ph.D. program. For 2021, U.S. News & World Report ranked LIS among the top 10 Library and Information Studies graduate programs in the country. The department is characterized by a research and learning culture that is inclusive, diverse and cohesive, international in scope but local in impact.