**Rutgers University, School of Communication & Information**

**Cultivating Equitable and Sustainable Ecosystems for Legacy Research Data**

IMLS National Leadership Grants for Libraries
Project Type: *Applied Research*

This proposed project 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."). It 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.

Rutgers University's School of Communication & Information, in partnership with Rutgers Libraries and Indiana University's (IU) Luddy School of Informatics, Computing & Engineering and IU Libraries, requests $194,634 to support a two-year Applied Research project to explore the existing ecosystems of legacy research data and 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** of determining what data to keep, why, how, and for how long.[1] The proposed project will **examine the 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.**

**PROJECT JUSTIFICATION:** Legacy data is data that has been collected in the past, yet it is often inaccessible to potential users[2,3]. Such data are abundant, and they are crucial in enabling open and innovative research, including computational data-driven modeling and forecasting and understanding long-term changes in our world. Astronomy, climate and biodiversity studies, archaeology, and many other domains depend on legacy research data. 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[4]. How can we (researchers, libraries, and society) capitalize on 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 preservation frameworks are imperative, and libraries are well positioned to do that by virtue of their stewardship mission and archival expertise.

Successful legacy data services depend on collaborative models in which researchers identify relevant data and work together with libraries and IT to curate and preserve them. Developing such collaborative models and sustaining them long-term, in turn, requires a better understanding of what working with and preserving legacy data means to various communities. While many examples of legacy data preservation and rescue initiatives exist, there has been no systematic study to date that synthesizes the lessons and insights from those separate initiatives and creates a map of the legacy research data ecosystem that can guide future efforts. The applied research project proposed here aims to fill this gap and **develop a systematic understanding of legacy research data efforts and the value of legacy data as it is perceived by various research, library, and IT communities.** The project will engage with the discourses and communities that are involved in legacy data preservation at both quantitative and qualitative levels and examine the socio-technical processes and impact and the role of equity, fairness, and justice in working with and curating legacy data.

**PROJECT WORK PLAN**: 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? Three integral themes will guide the study and its theoretical and methodological approach:

[1] Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review, 1*(1).
[2] Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library trends*, *57*(2), 280-299.
[3] Stahlman, G. R. (2020). Exploring the long tail of astronomy: A mixed-methods approach to searching for dark data (Doctoral dissertation, The University of Arizona).
[4] Griffin, E.R. (2015). When are old data new data? *GeoResJ*, 6, 92–97. DOI: 10.1016/j.grj.2015.02.004

**Rutgers University, School of Communication & Information**

**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. It will help identify existing patterns and opportunities for creating and maintaining such linkages.
**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, and the degrees to which institutions and individual researchers are willing to share their physical and intellectual resources. It will guide the development of research instruments in emphasizing collaboration and sharing, with a particular emphasis on inequities in access and use.
**Theme 3:** ***Engaging local communities*****.** Many disciplines have a history of colonialism while collecting data in locations of geographic interest. For example, indigenous and local communities have argued that astronomers exploit the lands upon which they place telescopes and collect data. Moreover, local and under-resourced researchers may not have access to digitized data as such access requires computing resources. Our research will pay particular attention to these aspects of legacy research data ecosystems and examine how to diversify management, access, and use of legacy data.

We will address these themes and questions by implementing a **mixed-method approach**. First, we will **construct a corpus** of representative publications and projects and annotate and analyze them using content analysis and scientometric methods, looking for references to individuals and institutions involved, processes, data types, formats, locations, and problems being addressed. Such a study will provide a systematic review of the legacy data preservation landscape and identify existing stakeholders, approaches, and models of sustainability and collaboration. Second, we will **conduct a survey of academic librarians and archivists** aimed at understanding current attitudes toward and activities around research legacy data. Third, we will engage with **three legacy data migration and preservation case study sites** for focused qualitative investigation into underlying processes, technologies, and broader socio-technical and communication ecosystems. Two preliminary cases have been identified for this project, and one additional case will be identified via the corpus analysis. All outcomes will be shared with broader research and data management communities for **peer feedback** before they are synthesized into a conceptual framework. Finally, our activities and findings will be translated into **open educational resources** that we will develop and pilot in LIS courses (such as Digital Curation at Rutgers and Management, Access, and Use of Big and Complex Data at Indiana University).

The cross-institutional leadership team will include co-PIs Stahlman (Rutgers) and Kouper (IU), who have extensive expertise in qualitative and mixed-method research and data curation. A graduate student will be recruited to assist with research activities and gain knowledge and skills about legacy data curation to be deployed on the ground as a future information professional.

**DIVERSITY PLAN**: The project will examine connections between local and indigenous knowledge and legacy data (Theme 3 above) 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 graduate student assistant from an underrepresented group will be recruited to assist with the project. All collaboration and dissemination efforts will prioritize under-resourced and underrepresented institutions and communities.

**RESULTS AND IMPACT**: The deliverables and intellectual outcomes of this proposal include: 1) **Understanding of the current approaches** to curating and preserving legacy data in academic libraries in the US; 2) **Value assessment of researchers' needs** for legacy data in several disciplines as well as impacts on local communities; 3) **A framework for legacy research data services** that enables further development of equitable and sustainable workflows and policies**,** 4) **A step-by-step guide** on curating specific legacy data, including contact with researchers, assessment, format migration, cleaning, and preservation; and 5) **Open curricular material** for training LIS 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.

**BUDGET SUMMARY**: The total budget for this 2-year project is $194,634. This includes $60,725 salary and fringe benefits for one student assistant at Rutgers plus one course release and one month of summer salary for PI Stahlman; and $13,700 travel for PIs' site visits and dissemination (F&A $56,672). This also includes a $63,537 subaward to Indiana University to cover $39,087 salary and fringe benefits and $1,000 supplies for co-PI Kouper (F&A $23,451).