Data Curation for Reproducibility (Data CuRe) Training Program Planning

Overview

The Odum Institute at the University of North Carolina, in collaboration with the Institution for Social and Policy Studies (ISPS) at Yale University and the Cornell Institute for Social and Economic Research (CISER) at Cornell University, request $49,890 to plan and develop an evidence-based training program focused on data curation for reproducibility. The integrity of the scientific record is a growing concern, amplified by recent reports of failed attempts to reproduce published findings. The cause of these failures varies in nature, but lack of access to well-documented data and executable analysis code underlying published results is a singular factor that renders impossible any effort to verify (and to extend, reinforce, and refine) research results. Funding agencies, journal editors, and research communities have become attuned to this fact and have issued policies and guidelines that require researchers to share their data and code as a means to promote reproducible research. In response, research institutions are mobilizing the expertise of librarians and archivists to provide support to researchers working to meet policy requirements and community expectations. A Planning Grant in the Curating Collections project category to support Continuing Education will support initial steps towards the development of a Data Curation for Reproducibility (Data CuRe) training program that will help fill gaps in the current skillsets of librarians and archivists adapting their roles to support the goals of reproducible research.

Statement of Need

As the number of data sharing policies increases, so too is the number of librarians and archivists working to acquire the specialized skills required to provide the data curation services designed to enable discovery, access, and reuse of research data. To support reproducible research, these data must meet a standard of quality as defined by the *replication standard*. The replication standard requires that "sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author".[1] Aligning data curation activities with the replication standard demands a workflow that includes file review, file normalization, metadata generation, assignment of persistent identification, data cleaning, and assembly of contextual documentation. While this list of standard data curation tasks alone is comprehensive, it does not necessarily guarantee that the data meet quality standards that allow for research reproducibility. What also must be included is a review of the computer code used to produce the analysis to ensure that the code is executable and generates results identical to those presented in the associated published article.

The project team represents organizations that have implemented data curation workflows, services, and tools in direct support of reproducible research. The Odum Institute has integrated data curation and code review workflows into the scholarly publication workflow to enforce journals' data replication policies; CISER has adopted standard protocols for research lifecycle-based data curation activities that enable researchers to produce high quality datasets that support reproducibility; and ISPS was instrumental in the conception of the data quality review framework, which informed their development of a tool that facilitates full data curation and

[1] King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452. https://doi.org/10.2307/420301

code review workflows. These are three of only a handful of organizations that have built capacity to implement data curation for reproducibility standards and activities. Many more will be needed to fulfill the research community's demand for access to datasets that meet the replication standard of quality.

The existing skillsets of librarians and archivists are known to lend themselves to the data curator role. However, the knowledge and skills necessary to perform full data curation for reproducibility goes beyond the current expectations and abilities of most librarians and archivists. Data curators tasked to review code must possess demonstrated knowledge of statistical software packages and research methodologies in order to understand the code, identify and diagnose errors, and enforce standards for interpretable, executable code. If libraries and archives are to be prepared to address heightened concerns about what has been referred to as the "reproducibility crisis," they will need to expand further their skillsets so that they can offer the data curation services the scientific community needs to help overcome this crisis. A continuing education program that provides opportunities for library and archives practitioners to augment their current expertise with the skills and knowledge necessary to perform intensive data curation tasks will allow them to respond to these pressing needs.

Planning Grant Activities

Planning Grant activities will enable the project team to take the first steps in developing a comprehensive Data CuRe training program. Grant activities will include the identification of the specific skills and knowledge required of librarian and archivists to perform data curation tasks that support research reproducibility—that are missing in current education programs. To do this, the project team will perform an environmental scan of the data curation practices of libraries and archives providing data curation services, review existing statistical computation training programs, and consider the data curation experiences of the project team's respective organizations. Based on the information gathered, the project team will produce a curricular framework that will serve as the foundation for a strategic plan for continued development of the Data CuRe training program.

Project Outcomes

1. Report describing the training needs of librarians and archivists appointed to perform data curation for reproducibility to be widely disseminated to the information professional community and other stakeholders in the research enterprise.
2. Curricular framework that outlines the essential components of a Data CuRe training program to include statistical computation, research methodologies, and principles of research reproducibility, to be widely disseminated to the information professional community and other stakeholders in the research enterprise to solicit feedback and input.
3. Strategic plan that articulates the project team's next steps towards the development of an evidence-based Data CuRe training program.

Budget

The project requests $49,890 to fund activities required to produce proposed outcomes. Grant funds will be used for PI salary and fringe ($5,566), travel to relevant professional events to disseminate project outputs ($4,500), subcontracts ($20,000) for collaborators' salaries, fringe, and travel. The proposed budget includes indirect costs calculated at the UNC F&A a rate of 36% ($12,624).