Implementing a Data Curation for Reproducibility (Data CuRe) Training Program

Overview

The Institution for Social and Policy Studies (ISPS) at Yale University, along with the Odum Institute at the University of North Carolina, and the Cornell Institute for Social and Economic Research (CISER) at Cornell University, request $208,610, to deploy the evidence-based Data Curation for Reproducibility (Data CuRe) training program. This proposal represents the implementation phase of the IMLS LB21 Planning Grant (RE-87-17-0074-17) that supported the curricular development and strategic planning for Data CuRe. Data CuRe was designed to provide critical training to library and archives practitioners who have been seeking opportunities to fill gaps in their current skillsets to provide support to members of the research community burdened with growing expectations for reproducible research practices. Central to these practices is the production of data and code that meet quality standards as defined by the FAIR Data Principles (https://www.force11.org/group/fairgroup/fairprinciples) and achieved by applying rigorous data curation and code review workflows prescribed by the Data Quality Framework (Peer & Green). A two-year Project Grant in the National Digital Infrastructures and Initiatives project category to support Continuing Education will enable deployment of the Data CuRe curriculum. In doing so, the proposed project will expand the community of practice around curating for reproducibility, which will become imperative as the research community continues to look to libraries and archives to provide the tools, services, and expertise to support latest norms in research practice.

Statement of Broad Need

As data sharing becomes normative practice in the research community, there is growing awareness that access to data alone – even well-curated data – is not sufficient to guarantee the reproducibility of published research findings. Computational reproducibility, the ability to recreate computational results from the data and code used by the original researcher, is a key requirement to enable researchers to reap the benefits of data sharing, but one that recent reports suggest is not being met. In addition to the underlying data, verifying findings to confirm the integrity of the scientific record and to build upon previous work to discover and develop new innovations also requires access to the analysis code used to produce reported results. The exhaustive laundry list of tasks that characterize the traditional data curation workflow that enables data access--file review and normalization, metadata generation, assignment of persistent IDs, data cleaning, and assembly of contextual documentation--falls short when research reproducibility is the ultimate goal (Peer et al., 2014). In order to curate for reproducibility, activities must include 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.

As determined during planning grant activities, library and archives professionals are becoming aware of the imminent demand for reproducible research support, but are but are not yet equipped to provide the data curation and code review services central to that support. An environmental scan of existing training programs and solicitation of community feedback revealed gaps in opportunities for library and archives practitioners to gain the necessary foundational knowledge of issues in reproducible research, computational skills to perform code review, and basic comprehension of statistical methods. The project team held several well-attended workshops to assess the effectiveness of the content and delivery of curation for reproducibility instruction, which indicated the need to further explain and contextualize fundamental concepts, simplify hands-on activities, and consider extending training duration. The promotion of Data CuRe attracted attention from other organizations with mechanisms in place to effectively deliver instruction to targeted audience that have welcomed strategic partnerships to complement and advance Data CuRe goals. These activities, along with the proven experiences of the project team, which represents organizations that have implemented data curation workflows, services, and tools in direct support of reproducible research, will inform the project design, activities, and desired outcomes of the CuRe program and help further the reproducibility of published research.

Project Design

Implementation of the Data CuRe training program developed during the planning phase will leverage partnerships with Project TIER to further enhance curricular modules with emerging standards for rigorous data curation, the Data Curation Network to promote Data CuRe to the targeted audience of librarians and archivists engaged in research data support, and Library Carpentry to deliver training using a proven mechanism for hands-on instruction. Grant activities will include:

1. Deployment of the Data CuRe training program to the intended audience of library and archives professionals who provide data support to the research community.
2. Wide dissemination of Data CuRe curriculum materials to the library and archives community and other stakeholders in the research enterprise for reuse and extension.
3. Formative and summative evaluation of the Data CuRe training program to inform ongoing improvement to address learner needs and emerging trends in research practice.

Diversity Plan

Yale University is committed to diversity as part of its mission to improve the world through outstanding research and scholarship, education, preservation, and practice by engaging in a free exchange of ideas in an ethical, interdependent, and diverse community. Likewise, diversity is reflected in the multicultural perspectives of project team members, who are committed to addressing diversity in the following ways: Involving academic libraries and archives at institutions that serve underrepresented groups (e.g., historically black colleges and universities) in the environmental scan and development of the curricular framework and strategic plan to ensure feasibility of implementation for a variety of institutions; considering different models for instructional delivery (e.g., modular online courses) to extend the reach of Data CuRe Training Program to a global audience, and to accommodate possible financial and time constraints of library and archives practitioners; attracting librarians and data curators of diverse professional and cultural backgrounds to participate in Data CuRe Training Program activities by disseminating information about the project to organizations that specifically serve underserved groups; and designing an accessible curricular framework to allow for individuals with disabilities to participate fully in all components of training and education programs.

Broad Impact

Beyond presenting opportunities to enhance the skillsets of library and archives professionals, Data CuRe offers broad impact that pushes the boundary of their role in sustaining the value of research assets and preserving the integrity of the scientific record as described below:

1. Expansion of the capacity of libraries and archives to promote and support the goals of reproducible research in response to scientific community needs
2. Increase in the quality of research outputs under the stewardship of libraries and archives that host data repositories
3. Growth of a community of practice engaged with the critical issues affecting the scientific community in order to define, develop, and deploy standards and best practice for data curation for reproducibility

Budget

The project requests $208,610 to fund activities required to produce proposed outcomes. Grant funds will be used for project team salaries and fringe ($107,465), travel to training and relevant professional events ($30,000), and web development ($6,500). The proposed budget includes indirect costs calculated at an off-campus rate of 26%.