Laura Bush 21st Century Librarian Program (LB21-FY19)
Institution for Social and Policy Studies, Yale University

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

Abstract

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 $ 241,214 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.

Funding agencies and journal publishers increasingly require sharing data to allow others to reproduce the results reported. Facilitating and enhancing reproducibility of published work is a key objectives of research data management, largely by making the methodology more transparent and helping verify research results over time. The imperative for reproducible research has been brought to the attention of libraries and archives, which are seeing an increase in demand for data curation support. However, the knowledge and skills necessary to perform data curation that includes a code review component—essential to research reproducibility—goes beyond the current expectations and abilities of most librarians and archivists.

A continuing education program targeting academic librarians and data archivists would offer opportunities for practitioners to augment current expertise with the skills and knowledge necessary to perform intensive data curation tasks and 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 and achieved by applying rigorous data curation and code review workflows prescribed by the Data Quality Framework. The Data CuRe Training Program will also expand the community of practice around curating for reproducibility, which is imperative as the research community continues to look to libraries and archives to provide the tools, services, and expertise to support latest norms and rigorous standards in research practice.

The primary activities of the Data CuRe Training Program project are as follows:

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

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 $241,214 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, 2012). 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. 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 and rigorous standards in research practice.

1. STATEMENT OF BROAD NEED

Funding agencies and journal publishers increasingly require sharing data to allow others to reproduce the results reported. Facilitating and enhancing reproducibility of published work is a key objectives of research data management (Vitale, 2016), largely by making the methodology more transparent and helping verify research results over time (Miller, 2018). The imperative for reproducible research has been brought to the attention of libraries and archives, which are seeing an increase in demand for data curation support. However, the knowledge and skills necessary to perform data curation that includes a code review component—essential to research reproducibility—goes beyond the current expectations and abilities of most librarians and archivists.

The goal of the Data CuRe Training Program is to help librarians and archivists better serve the research community. By combining the data management and curation skills that librarians and archivists master with the reproducibility skills that researchers are gradually adopting, the Program will help bridge between the two communities.

While academic libraries and data archives have been providing the systems and standards for making research materials publicly accessible, the datasets housed in repositories rarely meet the quality standards required by the scientific community. 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 (Stodden et al., 2013), 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 the code is executable and generates results identical to those presented in associated publications.

As determined during Planning Grant activities (see https://www.imls.gov/grants/awarded/re-87-17-0074-17), library and archives professionals are becoming aware of the imminent demand for reproducible research support, but are not yet equipped to provide the data curation and code review services central to that support. A recent study by Lisa Johnston and colleagues finds that while "code review" is a highly valued curation activity, it is among the services that is "not happening or not happening in a satisfactory way for a majority of our researchers" (2018, p.15).

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 scan helps identify the specific skills and knowledge required of academic librarians and archivists to perform data curation tasks that support research reproducibility and that are missing in current education and training programs. Expanding these skills presents an opportunity for libraries and archives to affirm their central role in academic research and to enrich the research enterprise by working alongside researchers and non-library information and data specialists (e.g., data scientists) who may have the computational and statistical skills but lack the curation and library perspective.

A continuing education program targeting academic librarians and data archivists would offer opportunities for practitioners to augment current expertise with the skills and knowledge necessary to perform intensive data curation tasks that support the research community's demand for high quality data and code.

The primary activities of the Data CuRe Training Program project are as follows:

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

This project will harness the experience of the CURE Consortium. Over the past 18 months, the project team has designed a CURE workshop based on team members' prior teaching experience, evaluation of training programs with similar learning objectives, and workforce needs of the library and archives community according to advisory group members and outputs from data curation initiatives (e.g., Data Curation Network). In addition, CURE-inspired workshops are now part of the offering at the Cornell Institute for Social and Economic Research (https://ciser.cornell.edu/transparent-and-reproducible-research/). The project team held several well-attended workshops to assess the effectiveness of the content and delivery of curation for reproducibility instruction. In addition to demonstrating that library and archive professionals are interested in the topic, the workshops indicated the need to further explain and contextualize fundamental concepts, simplify hands-on activities, and consider extending training duration.

The Data CuRe Training Program and its objectives build upon the work of other groups and individuals working to support the goals of reproducible science. The promotion of Data CuRe attracted attention from

several organizations in both the data curation space and the reproducibility space who have indicated a willingness to advise and share information to complement and advance Data CuRe goals, notably Library Carpentry, Project TIER, and the Data Curation Network. Please refer to supporting document 2, part 2 for more information about these projects and letters of commitment.

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.

2. PROJECT DESIGN

The project design represents steps in implementing a comprehensive Data CuRe Training Program, a continuing education program targeting academic librarians and data archivists.

The Data CuRe Training Program will build on a preliminary version of the curriculum developed with the support of an IMLS Planning Grant. The project team has taught the preliminary curriculum as a half-day workshop in conference and university settings. These workshops include both context and hands-on components. For context, lessons include the reproducibility crisis, curating for reproducibility (CURE) using the Data Quality Review (DQR) framework (Figure 1), and institutional models for practicing CURE. The hands-on component of the workshop includes a guided exercise in how to curate for reproducibility and perform a code review using DQR – including converting absolute file paths to relative file paths; checking code for presence of non-executable comments that document analysis processes; identifying packages required to execute code; executing code to ensure code is error-free; comparing code output to findings presented in paper; and more – and a demonstration of a new curation workflow tool.[1] Please refer to supporting document 2, part 1 for a sample workshop agenda based on the preliminary curriculum and a list of workshops conducted; workshop slides are available on OSF (https://osf.io/3wkex/).

Figure 1: Data Quality Review Framework (Peer et al., 2014)

The preliminary curriculum is based on the project team's extensive experience delivering services, developing tools, and overseeing infrastructure relating to curating for reproducibility. The Principals' location within university units practicing curation for reproducibility has the advantage of direct contact with users of these services and tools, which in turn informs both the services and the curriculum in an iterative process. The project team's vantage point led to important lessons learned in the Planning Grant that will be incorporated into the Data CuRe Training Program: To close the gap between LIS and researcher approaches to addressing the reproducibility crisis by including lessons on both data and code quality as a prerequisite to reproducibility (a LIS-focused approach) and on tools that support research transparency (a growing researcher-focused

[1] Version 1.0 of the software is currently used at ISPS, see https://isps.yale.edu/research/data/deposit/yard.

approach), and to seek collaboration with other groups working to enable researchers to produce and share data and code that meet standards for reproducibility.

The preliminary curriculum developed during the planning phase will inform a strategic plan for continued development and implementation of the Data CuRe Training Program. The main tenets of this plan are, a). adapting the Data CuRe Training curriculum to the Carpentries workshop model and, b). soliciting continuous feedback on project outputs throughout the project period. The Carpentries model establishes a framework for structuring and implementing participatory hands-on instruction that includes a computation component. Founded in 1998, the Carpentries has worked to fulfill its mission to "build global capacity in essential data and computational skills for conducting efficient, open, and reproducible research." For over 20 years, the Carpentries has refined its evidence-based pedagogy through a program of continuous assessment that evaluates the impact of Carpentries on the individuals and communities it serves. The results presented in the 2018 Carpentries Long-Term Impact Survey report (Jordan, 2018) showed that 77% of learners reported having greater confidence in the tools they learned at a Carpentries workshop six months after their attendance. To date, Carpentries has built a growing community of 1,640 instructors who have taught 1,703 workshops to 39,000 learners (http://static.carpentries.org/volunteer/). This community now includes information and library professionals with the addition of Library Carpentry to the existing Software Carpentry and Data Carpentry Lesson Programs. Currently, Library Carpentry does not include lessons about curating for reproducibility (https://librarycarpentry.org/lessons/). Having a common goal of providing opportunities to librarians and other data support professionals to acquire the data and software skills necessary to support reproducible research, the use of Library Carpentry as the primary deployment mode for CuRe training is both logical and fortuitous (https://docs.carpentries.org/topic_folders/governance/bylaws.html).

Project activities will be supported by the data curation and scientific community from whom the project team will solicit feedback on project outputs throughout the project period. The implementation of the Data CuRe Training Program will leverage relationships with Library Carpentry (https://librarycarpentry.org/) to deliver training using a proven mechanism for hands-on instruction, Project TIER (https://www.projecttier.org/) to further enhance curricular modules with emerging standards for rigorous data curation, and the Data Curation Network (https://datacurationnetwork.org/) to share curricula and experiences from developing specialized data curator training the targeted audience of librarians and archivists engaged in research data support. In addition, the project team will make use of other methods to solicit community input, similar to the interactive poster presented at IASSIST 2018 (please refer to supporting document 2, part 1).

Project Activities

To achieve its objectives, the project will consist of three primary activities: Implementation of the Data CuRe Training Program, outreach and dissemination, and community engagement.

1. ***Implementation of the Data CuRe Training Program***

   The Data CuRe curricular framework developed in the planning phase is informed by the project team's experience conducting workshops using the Data CuRe materials, conclusions from the environmental scan, and feedback from the Advisory Group. The implementation of the Data CuRe Training Program will involve refinement of the Data CuRe curriculum, development of additional materials, evaluation, and a phased deployment.

   The general outline of the curriculum is as follows:

a. What is CURE. This module will include two topics: Reproducibility and scientific research and the role of curation in reproducibility. The first topic will provide background on the scientific method and reproducibility, the reproducibility crisis, and approaches to improving reproducibility in the scientific community. The second topic will focus on the role of curation in reproducibility and the nexus of curation and reproducibility and introduce CURE standards and the Data Quality Review framework.
b. How to CURE. This module also includes more hands-on topics: Working with code and Implementing the Data Quality Review framework. The first topic will cover types of scripts used in the research lifecycle, types of software for using code, common issues with code in terms of reproducibility, and common issues with code in terms of digital preservation. The second topic will be dedicated to implementing the Data Quality Review framework and include code review as an essential skill for CURE, including curation activities such as maintaining, storing, and ensuring accessibility of code, approaches to code review, tools for code review, and step-by-step how to conduct code review.

Please refer to supporting document 2, part 2 for preliminary curriculum planning, 2019-2021.

The deployment of the Training Program will involve two phases. In Phase 1, the project team will pilot the curriculum developed in the Planning Grant and further refine the materials in line with Library Carpentry's guidelines, in consultation with Project TIER, and in coordination with the Data Curation Network. The project team plans to hold three workshops in Phase 1. This phase will include user testing and the application of instructional design principles. Evaluation of this phase will inform the next phase of the project (see section, "Project Evaluation"). In Phase 2, the project team will finalize the curriculum and fully integrate with Library Carpentry to achieve teaching at scale. In this phase, the Data CuRe Training Program will be a fully branded Carpentry lesson. The project team plans to teach the lesson via Library Carpentry three times in Phase 2. Please refer to the Schedule of Completion document for detailed information and timelines for individual project activities.

The implementation of the Training Program will consist of two modes of delivery consistent with Library Carpentry guidelines: In-class and online. Carpentry lessons are meant to be taught at official Carpentry workshops by a Carpentries-certified instructor, and the project team will work within that framework to develop the lessons and to leverage Library Carpentry's proven mechanism for hands-on instruction. For in-class training, the project team will leverage opportunities to recruit participants who are already attending relevant conferences and meetings, especially during phase1, adding official Library Carpentry workshops in phase 2. In accordance with Carpentry guidelines, the Training Program will be implemented such that lessons will be available online and can be used in other types of workshops, or added to graduate level courses, bootcamps, and other types of training.

The project team is designing the Data CuRe Training Program to have direct impact on the LIS community by making it cost efficient and widely accessible. Based on the curriculum developed in the Planning Grant, as well as on Library Carpentry guidelines, the Data CuRe curriculum will likely span a half a day. In-class training will take place at, and outreach to participants will be organized through, professional conferences, both domestic and international, and workshops organized via Library Carpentry and hosted by academic institutions across the United States. Effort will be made to reach institutions that serve underrepresented groups (e.g., historically black colleges and universities) and to attract librarians and data curators of diverse professional and cultural backgrounds. The Library Carpentry model seeks to keep the cost of hosting

workshops to a minimum (https://librarycarpentry.org/get_involved/) and there is typically no cost to participants. Based on the project team's experience during the Planning Grant, in-class training is estimated at about 25 participants, which is also in line with Library Carpentry's guidelines for helper-to-learners ratio of 1:8, given the project team consists of three Principals. Based on our experience during the Planning Grant, the project team estimates reaching a minimum of 150 LIS professionals directly. Further, the Library Carpentry model of "train the trainer" puts in place a mechanism to maximize the likelihood that future LIS professionals will also benefit from this curriculum. In addition, and to broaden the reach of the Data CuRe curriculum, the materials will be made available online and can be incorporated into other trainings or for self-directed learning.

2. Outreach and dissemination

To ensure the broadest reach and community engagement, the project team will employ an outreach and dissemination strategy that promotes full awareness of CuRe training opportunities and open access to curriculum materials through several venues.

Project website. All project outputs will be made openly accessible via the CURE Consortium website, which will contain all CuRe curriculum materials and current information on training events.

Library Carpentry website. The CuRe Training Program will be included as a Library Carpentry lesson, open to all library and archives practitioners with a basic level of experience in data management or curation.

Social media. News of CuRe training events will be announced via social media platforms including listservs and Slack workplaces that host LIS communities and other relevant stakeholders, as well as Twitter. Announcements will also be sent directly to other organizations for distribution to their audiences.

Professional conferences. Project outputs will be shared at professional conferences that are well-attended by library and archives practitioners who provide or plan to provide data services to the research community. These conferences will include:

- Research Data Alliance (RDA) Plenary Meeting
- FORCE11 Annual Conference
- International Digital Curation Conference (IDCC)
- Research Data Access & Preservation (RDAP) Summit
- International Association for Social Science Information and Technology (IASSIST) Annual Conference

Publication. The project team will produce manuscripts to be submitted for publication in open access scholarly journals that serve LIS audiences and/or research communities more broadly. All project reports and other articles will be made available on preprint servers.

The project team will encourage reuse and extension of project outputs by assigning materials a Creative Commons Attribution license. This license allows individuals to copy, redistribute, transform, and build upon the materials with no restrictions beyond giving credit to the CuRe project.

3. Community engagement

The project team will engage with established organizations in the research data space, the Research Data Alliance and FORCE11, that have a proven record of building communities around a topic and delivering

concrete solutions. These organizations are appropriate vehicles for such engagement because they a). have established processes for creating interest and working groups in the relevant communities and, b). lack groups that are working on curating for reproducibility.

The project team will engage in coordinated efforts with the Data Curation Network, Project Tier, Library Carpentry and other groups such as, Whole Tale and the Inter-university Consortium for Political and Social Research (ICPSR). In doing so, each group can pursue its goals more effectively by sharing resources and expertise to avoid redundancy in efforts, reconciling divergent uses of relevant terms and definitions for reproducibility concepts, addressing issues most relevant to a diverse community of stakeholders, and increasing public awareness of our efforts.

The project team intends on holding events at disciplinary and professional association meetings in order to interact with researchers directly, for example, the American Political Science Association (APSA), American Economic Association (AEA), American Sociological Association (ASA), American Association for Public Opinion Research (AAPOR), the American Psychological Association (APA), and the Population Association of America (PAA). As the research community is a benefactor of our efforts, direct communication about the Data CuRe project will provide an continuing feedback loop and help narrow any gaps in understanding and language and ensure that content development in on track.

Project activities will also be informed by feedback solicited from the data curation and research communities. The project team will engage the community during presentations of project progress and outputs at professional conferences such as the International Association for Social Science Information Service and Technology (IASSIST), the International Digital Curation Conference (IDCC), and the ASIS&T Research Data Access and Preservation (RDAP) Summit. Each year, hundreds of information professionals gather to discuss issues relevant to the data curation community.

The project team will also leverage the expertise of the CURE Consortium advisory group for feedback and advice. Please refer to supporting document 2, part 2 for letters of support. The advisory group represents leaders in data curation practice and initiatives supporting research reproducibility, with positions in academic libraries, domain-based data repositories, LIS graduate programs, and open science advocacy groups. Members of the CURE Consortium advisory group are:

- **Jacob Carlson**
  Director of Deep Blue Repository and Research Data Services
  University of Michigan Library
- **Colin Elman**
  Associate Professor of Political Science and Director of the Qualitative Digital Repository
  Syracuse University
- **Ann Green**
  Consultant and Strategic Analyst
- **Lisa Johnston**
  Director, Data Repository for the University of Minnesota (DRUM)
  Principal Investigator, Data Curation Network
- **Jared Lyle**
  Director, Metadata & Preservation
  Inter-university Consortium for Political and Social Research, University of Michigan
  Director, DDI Alliance

- **Victoria Stodden**
  Associate Professor
  School of Information Sciences, University of Illinois at Urbana-Champaign

Project Evaluation

Building on assessment of the CuRe curriculum during the planning phase of the project, the project team will conduct a formal program evaluation that assesses progress toward and achievement of project aims during the proposed training program implementation phase. The program evaluation will be both formative and summative in nature, with formative assessment offering insights that will enable the project team to make improvements to components of the training program during the iterative enhancement process, and summative assessment measuring the success of the training program pilot to achieve project goals. As such, evaluations will consist of several assessment methods to measure project success based on the evaluation methods as follows.

Formative evaluation. The formative evaluation will measure the degree to which CuRe training content and delivery successfully supports learners' achievement of learning objectives. To do this, the project team will administer pre- and post-workshop surveys, which will include IMLS learning measure statements, to CuRe training participants to collect information on individual learners' knowledge and skills related to reproducible research before and after workshop events. The project team will also solicit feedback from Advisory Group members at points before, during, and after deployment of the CuRe training program.

Summative evaluation. The summative evaluation will measure project success in achieving its primary objectives. Successful deployment of the DataCuRe Training Program to the intended LIS practitioner audience will be based on the number and characteristics of workshop attendees. The project team will also track reuse of CuRe training materials to determine success in disseminating curriculum materials to both the LIS community and other research stakeholders.

These evaluations will be included in reports to be shared with stakeholders to contribute to broader community efforts in the development and delivery of effective training and education in reproducible research.

Project resources

The project requests $241,214 over two years to fund activities required to produce the proposed outcomes. Grant funds will be used for the PI's salary and fringe and travel to training, meetings, and relevant professional events ($66, 343), subcontracts for co-PIs' salaries and fringe and travel ($120,052), travel for a convening of the advisory group members ($13,000), and web development ($6,500). The proposed budget amount is inclusive of the Yale Facilities and Administration Cost Rate of 26.00% for off-campus instruction applied to direct costs. For full budget information, please refer to the Budget Form, Budget Justification, and supporting document 3.

Project Personnel

The proposed project benefits from the strength of the collaboration and expertise of established data curation professionals working in well-known research institutions. Each team member possesses practical experience in the execution of data curation for reproducibility workflows and is contributing to the formation of data curation standards and best practices in this emerging area of archival practice.

The project will be led by Limor Peer, PhD (PI), Associate Director for Research at the Institution for Social and Policy Studies (ISPS) at Yale University, who will oversee the administration of the project. Limor Peer was the lead author of the article that first published the Data Quality Review framework that introduced code review into the data curation workflow (Peer et al., 2014). Limor led the creation of a specialized research data repository for ISPS scientists to support reproducibility of their research. All datasets housed in the repository have undergone the rigorous data curation process of Data Quality Review. Limor is also leading a project to develop YARD, the Yale Application for Research Data, a tool that structures the curation workflow for reviewing and enhancing research outputs before publication (Peer & Wykstra, 2016). Limor sits on the board of the Roper Center for Public Opinion Research and serves on a number of advisory and task force groups working on data curation and research transparency. At Yale she has been involved in various campus-wide strategic initiatives around research data policy and services at Yale, including two years at the Office of the Vice Provost for Research as a Research Data Specialist (2016-2018).

Thu-Mai Christian (co-Investigator) is Assistant Director for Archives at the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (UNC). Thu-Mai will serve as co-I who will lend her expertise in data curation standards and workflows to help develop the curricular framework for the Data CuRe Training Program. Thu-Mai has led the development of a data curation and verification service for academic journals that supports implementation and enforcement of the most rigorous of journal data policies. Working with journal editors, this service integrates the manuscript publication and data curation for reproducibility workflows to ensure that data underlying results in the journal publication meet the highest standard of quality (Christian et al., 2018). Thu-Mai was also the co-principal investigator for the IMLS-funded Curating Research Assets and Data Using Lifecycle Education (CRADLE) project (RE-06-13-0052), which produced a massive open online course in research data management and sharing that has served over 17,500 visitors and 4,000 active learners since its launch in February 2016.

Florio Arguillas (PhD, co-Investigator) is a Research Associate at the Cornell Institute for Social and Economic Research (CISER) at Cornell University. Florio will serve as co-I who will use his experience in providing data curation for reproducibility services to help identify the gaps in the education needs of information professionals tasked with data quality and code review responsibilities. At Cornell, Florio assists campus researchers in qualitative and quantitative data management and processing, assessing and mitigating disclosure risks, and use of statistical software packages. Florio has designed and established the Data Curation and Reproduction of Results Service, or R^2, which allows researchers to submit their data and code to CISER prior to manuscript submission for appraisal, curation, and replication by CISER data curation experts. Florio is internationally recognized as an expert in increasing the transparency and reproducibility of scholarly research and is a leading proponent of providing data curation and reproduction-of-results services to simplify replication and accelerate the scholarly process. Florio has also significantly contributed to simplification of data delivery, by developing software widely used by researchers and by organizations such as the Roper Center for Public Opinion Research. William Block, PhD, Director of the Cornell Institute for Social and Economic Research (CISER) at Cornell University will oversee the project as the Cornell PI.

The project will also include a Web Developer, TBD. The Web Developer will support the project by upgrading the CURE Consortium website (http://cure.web.unc.edu/) and by meeting the requirements of Library Carpentry for setting up a new lesson.

Project Timeline

The project design is based on a two-year award period of performance starting September 1, 2019 and ending August 31, 2021. Project activities in Year 1 will begin with a project team meeting, a focus on the refinement of the Data CuRe framework and the development of additional materials, and a Phase 1 deployment. Year 2 will include phase 2 deployment. Assessment and engagement with the data curation and research communities

will take place throughout the project period to enable incorporation of feedback into final project deliverables to be disseminated at the end of the project period. Please refer to the Schedule of Completion document for detailed information and timelines for individual project activities.

Project Dissemination

The project team will communicate the progress and products of the project through a variety of platforms including the CURE Consortium website (http://cure.web.unc.edu/), an OSF project page (https://osf.io/3wkex/), relevant listservs, social media, and professional conferences. The project will also increase awareness of Data CuRe Training Program plans by soliciting their feedback on iterations of curricular framework and strategic plan drafts. Project deliverables will be made accessible via the CURE Consortium website and stored in the UNC Dataverse repository for long-term preservation and access. A CC BY-SA 4.0 Creative Commons license will be applied to all products to encourage broad dissemination and reuse of the materials. Please refer to the Project Activities section in this document for detailed information.

3. 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.

4. BROAD IMPACT

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

a. Expanding the workforce capacity of librarians and archivist with the computational skills to promote and support the goals of reproducible research in response to scientific community needs.
b. Increasing the quality of research outputs under the stewardship of libraries and archives that host data repositories.
c. Growing 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.

The Data CuRe Training Project will expand the capacity of current and future library and archives professionals to support reproducible research, increase the quality of research outputs, and propel the role of libraries and archives as key research partners in sustaining the value of research assets and preserving the integrity of the scientific record.

Laura Bush 21st Century Librarian Program (LB21-FY19)
Institution for Social and Policy Studies, Yale University

Data Curation for Reproducibility (Data CuRe) Training Program Planning
SCHEDULE OF COMPLETION
The following table outlines major project activities as defined in the proposal narrative, to be completed during the project period starting July 1, 2019 and ending June 30, 2021.

Year 1

	2019				2020							
	Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
1. Implementation of the Data CuRe Training Program												
Curriculum & Library Carpentry meeting												
Refinement of the Data CuRe framework												
Adapting to Library Carpentry structure												
Deploy Data CuRe curriculum, phase 1												
User testing and evaluation, phase 1												
Deploy Data CuRe curriculum, phase 2												
User testing and evaluation, phase 2												
2. Outreach and dissemination												
Web development												
Disseminate project outputs via professional conferences, webinars												
Disseminate project outputs on project website and data repository												
3. Community engagement												
Expand CURE Consortium membership												
Engage with working groups within existing organizations												

Year 2

	2020				2021							
	Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
4. Implementation of the Data CuRe Training Program												
Curriculum & Library Carpentry meeting												
Refinement of the Data CuRe framework												
Adapting to Library Carpentry structure												
Deploy Data CuRe curriculum, phase 1												
User testing and evaluation, phase 1												
Deploy Data CuRe curriculum, phase 2												
User testing and evaluation, phase 2												
5. Outreach and dissemination												
Web development												
Disseminate project outputs via professional conferences, webinars												
Disseminate project outputs on project website and data repository												
6. Community engagement												
Expand CURE Consortium membership												
Engage with working groups within existing organizations												

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (e.g., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

All applications must include a Digital Product Form.

☐ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

All materials produced by the proposed project will be assigned a CC BY-SA 4.0 Creative Commons license that will allow others to copy, adapt, and share the materials with only the minimal requirements of giving appropriate credit to the project team and distributing derivative versions of the materials under the same CC BY-SA 4.0 license. To encourage broad dissemination and reuse of the materials, no other restrictions will be applied.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The project team will assert its right of first use of the materials produced by the project. All materials will be made publicly available for wide dissemination immediately after the project team has had the opportunity to make use of the materials for presentation and publication, and prior to the end of award period of performance.

A. 3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

The proposed project does not include the creation of products that raise privacy concerns.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

The project will produce reports, white papers, articles, presentation materials, and other documents containing details of project activities and findings. Final documents will be made available in standard .pdf file formats. Materials produced as part of Library Carpentry adhere to Library Carpentry lesson development requirements, including writing lessons in Markdown, and copies will be also made available in standard .pdf file format.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

Project materials will be produced using widely adopted business application software packages such as Microsoft Office (Word, Excel, PowerPoint). For long-term preservation of documents, final versions of files will be stored in Adobe .pdf format.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

XLSX
PDF
PPT
PDF (PDF/A for final versions for preservation purposes)
HTML

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

The project will solicit feedback on project activities and the quality of project materials from members of the advisory group.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

The project team will employ standard file sharing systems (e.g., Box, OneDrive) maintained by its respective institutions to store and manage project materials during the award period of performance. These systems include provisions for collaborative authoring, file versioning, and storage backups. Final products will be made publicly available via the project website and the project Open Science Framework site, and archived in the UNC Dataverse repository for long-term preservation and access during and after the award period of performance. Additionally, per Library Carpentry protocol, lessons will be stored in Git repositories on GitHub. A GitHub-Zenodo integration allows for GitHub repository archiving.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

Metadata will be produced using the UNC Dataverse repository interface. The Dataverse system uses metadata forms with fields corresponding to standard descriptive metadata elements based on the DataCite Metadata Schema. This metadata is stored and preserved alongside the archived materials.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

The file sharing systems to be used during the award period of performance captures technical metadata that records file information (e.g., file size, file version) and administrative metadata that tracks file activity (e.g., file uploads, file access, file modifications). Metadata describing final archived materials will be stored alongside the materials in the UNC Dataverse for long-term preservation and access during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

Standardized descriptive metadata stored in the UNC Dataverse alongside the final archived project materials will be discoverable through the UNC Dataverse search interface and the search interface of other repositories, including the Harvard Dataverse, that harvest UNC Dataverse metadata records through OAI-PMH.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

Project materials will be made publicly available on the project website as well as via the UNC Dataverse. The availability of materials will be announced on social media platforms, listservs, at professional conferences and meetings, and other available information outlets.

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

https://cure.web.unc.edu/
https://osf.io/3wkex/

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

Not applicable

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

Not applicable

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

Not applicable

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

Not applicable

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

Not applicable

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

Not applicable

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

Not applicable

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

Not applicable

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

Not applicable

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

Not applicable

URL:

Not applicable

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

Not applicable

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

Not applicable

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

Not applicable

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

Not applicable

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Not applicable

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

Not applicable

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Not applicable

A.8 Identify where you will deposit the dataset(s):

Name of repository:

Not applicable

URL:

Not applicable

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

Not applicable