Data Curation for Reproducibility (Data CuRe) Training Program Planning

ABSTRACT

The Odum Institute at the University of North Carolina, in collaboration with the Institution for Social and Policy Studies at Yale University and the Cornell Institute for Social and Economic Research at Cornell University, request $49,820 to fund planning activities for the development of an evidence-based training program focused on data curation for reproducibility (Data CuRe).

As the number of funding agencies, journals, and research institutions issuing data sharing policies increases, so too has 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. Advances in technology are transforming normative scientific practices now being defined by computational methods that rely heavily on data to arrive at results. The published article alone is no longer sufficient to disseminate scientific discoveries; data and code used by the author to produce results are vital to the understanding, validation, and extension of those results presented in the published article.

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. 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 rigorous 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 that support the research community's demand for high quality data and code.

The Data CuRe Training Program planning project will achieve the following objectives during the one-year award period:

- Analyze existing data curation for reproducibility workflows to inventory the skills and knowledge required of librarians and archivists to perform data curation for reproducibility.
- Examine present data curation education and training programs to identify gaps in existing opportunities for librarians and archivists to gain data curation for reproducibility skills.
- Review current training programs in statistical methodologies and computation to discover instructional methods that can be applied to a Data CuRe training curriculum.
- Produce an evidence-based curricular framework that addresses the education and training needs of librarians and archivists engaged in data curation for reproducibility activities.
- Engage the library and archives community in defining the role of the academic library and data archives within the emergent landscape of computational, data-driven research.
- Develop a strategic plan for the impending implementation of the curricular framework.

The Data CuRe Training Program planning grant represents the first steps toward the development and implementation of an evidence-based continuing education program that will enable librarians and archivists to continue to provide comprehensive data curation services.

Laura Bush 21st Century Librarian Program (LB21-FY17-2)
Odum Institute, University of North Carolina at Chapel Hill

Data Curation for Reproducibility (Data CuRe) Training Program Planning
NARRATIVE

> ***"…this new data dominant era brings new challenges for the scientists and they will need the skills and technologies both of computer scientists and of the library community to manage, search, and curate these new data resources. Libraries will not be immune from change in this new world of research"*** (Hey & Hey, 2006, p. 516).

1. STATEMENT OF NEED

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 $49,820 to fund planning activities for the development of an evidence-based training program focused on data curation for reproducibility (Data CuRe).

As the number of funding agencies, journals, and research institutions issuing data sharing policies increases, so too has 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. Advances in technology are transforming normative scientific practices now being defined by computational methods that rely heavily on data to arrive at results (Yale Roundtable, 2010). The published article alone is no longer sufficient to disseminate scientific discoveries. The availability of data and methods used to produce research results are vital to the understanding, validation, and extension of those results presented in the published article. As such, the scientific community has begun to recognize data as first-class research outputs in their own right (Callaghan et al., 2012). Because it is the mission of the library and archive to provide access to scholarship, libraries and archives have been acquiring the tools and skills to store, curate, and provide access to data so that their collections continue to represent the complete range of research outputs (Hey & Hey, 2006). In doing so, libraries and archives have been central to the development of the data curation standards and best practices that support data access (National Research Council, 2015).

Not surprisingly, the majority of recommendations to scientists, funding agencies, and publishers to increase research transparency and reproducibility through data sharing rely heavily on systems and standards established by the digital libraries and archives (Yale Roundtable, 2010). The FAIR Data Principles that present guidance for making data Findable, Accessible, Interoperable, and Reusable emphasize the use of standardized metadata specifications, controlled vocabularies, persistent and unique identification, and other practices (FORCE11, 2016)—practices that have their roots in digital libraries and archives. The Transparency and Openness Promotion Guidelines developed by the Center for Open Science make repository deposit of data and code a condition of policies promoting the most rigorous of reproducibility standards (Nosek et al., 2015). The recently published National Academies of Science (2017) report, *Fostering Integrity in Research*, makes an explicit recommendation to journals to implement open data policies, with the data repository the preferred method for accessing data underlying published findings. Indeed, libraries and archives have been experiencing increased demand from researchers for support as pressures mount to share data (Reilly, 2012).

Even so, the integrity of the scientific record continues to draw scrutiny from individuals questioning the reproducibility of published research findings. Despite the availability of platforms for sharing data, the quality of datasets housed in repositories do not necessarily meet the quality standards that allow for interpretation, verification, and reuse (Iqbal et al., 2016; Ioannidis et al., 2009; Peer, Green, & Stephenson, 2014). To support reproducible research, data must meet a standard of quality as defined by the *replication standard*, which 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" (King, 1995, p. 444). 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 is alone exhaustive, it cannot guarantee that the data meet quality standards that allow for research reproducibility. What also must be included in the laundry list of data curation activities is a review of the computer code used to produce the analysis (Peer, Green, & Stephenson, 2014). Code review ensures that the analysis code is executable and generates results from the data that are identical to those presented in the associated published article. This process is referred to as *computational reproducibility* (Stodden, Bailey, & Borwein, 2013), which is an essential criterion for data quality and fundamental to reproducible science.

The knowledge and skills necessary to perform rigorous data curation that includes verification of computational reproducibility—i.e., data curation for reproducibility—goes beyond the current expectations and abilities of most librarians and archivists. Reviewing code requires demonstrated knowledge of statistical software packages and research methodologies in order to interpret the code, identify and diagnose errors, and enforce standards for interpretable, executable code. If libraries and archives are to maintain their support for the research community and its imperative for reproducible research, librarians and archivists will need to expand further their skillsets. While traditionally beyond the purview of library and archive work, 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, as well as other information and data specialists working in research support units, 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 proposed planning project represents initial steps toward the development of a comprehensive continuing education program that will fill gaps in the current knowledge and skillsets of library and archives professionals engaged in data curation activities. The primary objectives of the Data CuRe Training Program planning project are as follows:

- Analyze existing data curation for reproducibility workflows to specify the data curation tasks that support research reproducibility and to inventory the specific skills and knowledge required of librarians and archivists to perform these tasks.

- Examine present data curation education and training programs to identify gaps in existing opportunities for librarians and archivists to gain the necessary skills to perform data curation for reproducibility.
- Review current training programs that provide instruction in statistical methodologies and computation to discover instructional methods that can be applied to data curation for reproducibility training curricula.
- Produce an evidence-based curricular framework that addresses the education and training needs of librarians and archivists engaged in data curation for reproducibility activities.
- Develop a strategic plan for the implementation of the curricular framework to include information on training activities and deliverables, milestones and projected timelines, resource needs (i.e., personnel, funding), and mechanisms for training program evaluation.
- Engage the library and archives community in defining the role of the academic library and data archives within the emergent landscape of computational, data-driven research.

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 Data Curation Network is developing a model for networked data curation services that enables individual academic libraries to harness the collective expertise and resources of network partners (Data Curation Network, 2016). The Opening Reproducible Research project is designing the Executable Research Compendium (ERC), a technological solution for supporting computational reproducibility informed by the unique perspectives and requirements of individual stakeholders involved in the scholarly communications processes (Nüst et al., 2017). The Project TIER (Teaching Integrity in Empirical Research) Protocol includes a specification that outlines the component parts of a complete, reproducible data package (Project TIER, 2016). Library Carpentry, an offshoot of Software Carpentry, offers software skills training adapted for a librarian audience (Baker et al., 2016). Victoria Stodden, who has been a leading figure in defining requirements for computational reproducibility, was the lead author of a set of recommendations that prescribe specific practices that yield reproducible research (Stodden et al., 2016).

While these initiatives have made significant contributions to reproducible research through the development of service models, tools, standards, and best practice recommendations for data access and research transparency, none directly prescribes the data curation workflows and tasks necessary to sustain the goals of reproducible science, nor the skillsets required to execute these workflows and tasks.

The project team represents three organizations that have been working to develop effective data curation workflows, services, and tools in direct support of reproducible research practices. 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 workflow tool that facilitates full data curation and code review workflows. These are three of only a small handful of organizations that have built capacity to implement data curation for reproducibility standards.

Many more will be needed to fulfill the research community's demand for access to datasets that meet the replication standard of quality.

Recognizing this fact, members of the project team recently formed the Curating for Reproducibility (CURE) Consortium, a group of academic institutions that have adopted or support data quality review, a framework that includes research data curation and code review (Peer, Green, & Stephenson, 2014). The CURE Consortium is committed to building a community of practice around the establishment of standards and best practices for data curation for reproducibility, sharing information on implementation of these standards and best practices, and promoting the philosophy of data quality review foundational to reproducible research. An initial aim of the CURE Consortium is to build capacity in data curation for reproducibility through education and skills training for academic librarians and data curators.

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 training program that will help fill gaps in the current skillsets of academic librarians and data archivists adapting their roles to support the goals of reproducible research.

2. PROJECT DESIGN

The project design represents initial steps in developing a comprehensive Data CuRe training program. Grant activities will include the identification of the specific skills and knowledge required of academic librarians and archivists to perform data curation tasks that support research reproducibility—that are missing in current education and training programs. 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 and implementation of the Data CuRe Training Program. 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.

Project Activities

To achieve its objectives, the project will consist of four primary activities: 1) environmental scan, 2) curricular framework development, 3) strategic planning, and 4) community engagement.

1. ***Environmental scan***
   The project team will perform an environmental scan to identify the knowledge and skillsets required of librarians and archivists to perform data curation for reproducibility tasks. The environmental scan will include an investigation of existing data curation for reproducibility workflows at a granular level that will allow for mapping of tasks to required skills. In addition, the project team will review existing data curation education programs including course syllabi and training materials to determine gaps in instruction that may be filled by continuing education opportunities. An examination of education and training programs in statistical methods and computation will also be included in order to discover instructional techniques that may be applied to the code review instruction components of the Data CuRe

curriculum. Finally, a survey of tools available to assist with code review will also be performed.

2. ***Curricular framework development***
   Based on information gathered from the environmental scan, the project team will develop a Data CuRe curricular framework that outlines learning objectives, subject matter and content, methods for content delivery, infrastructure and resource requirements, and learning evaluation approaches. The framework will map learning objectives to standards for data curation for reproducibility while also addressing the diversity of learners in terms of professional and cultural background.

3. ***Strategic planning***
   The project team will formulate a strategic plan for the implementation of the Data CuRe curricular framework. The strategic plan will summarize the next steps for producing and delivering the Data CuRe Training Program. Included in the plan will be:

   - a reiteration of the purpose and goals of the Data CuRe Training Program as it supports reproducible research through data quality review framework-based data curation;
   - identification of target audiences and other stakeholders;
   - timelines for Data CuRe course development and delivery;
   - a program evaluation strategy; and
   - strategies for sustainability through grant funding opportunities, contracts, and institutional support.

4. ***Community engagement***
   Project activities will be informed by feedback solicited from the data curation and research communities. The project team will engage the community during presentations of planning project outputs at professional conferences such as the International Association for Social Science Information Services and Technology (IASSIST) 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. The advisory group represents leaders in data curation practice and initiatives supporting research reproducibility. Members of the CURE Consortium advisory group are:

   - **Jacob Carlson**
     Research Data Services Manager
     University of Michigan Libraries
   - **Colin Elman**
     Associate Professor of Political Science and Director of the Qualitative Digital Repository
     Syracuse University
   - **Ann Green**
     Consultant and Strategic Analyst

- **William G. Jacoby**
  Professor of Political Science and Editor, *American Journal of Political Science*
  Michigan State University
- **Sarah Jones**
  Senior Institutional Support Officer
  Digital Curation Centre, University of Edinburgh
- **Jared Lyle**
  Director of Curation Services
  Inter-university Consortium for Political and Social Research, University of Michigan
- **Robin Rice**
  Data Librarian
  University of Edinburgh
- **Jeffrey Spies**
  Co-founder and CTO
  Center for Open Science
- **Victoria Stodden**
  Associate Professor
  School of Information Sciences, University of Illinois at Urbana-Champaign

Project Evaluation

Project deliverables will undergo continuous and iterative evaluation to assess the project success and areas for improvement. The project team will collect and synthesize feedback received from the data curation and research community and the CURE Consortium advisory board on drafts of the CuRe Training Program curricular framework and strategic plan for CuRe Training Program implementation. This feedback will be documented in reports, which will help inform future steps towards establishing a comprehensive training program in data curation for reproducibility. These reports will be widely disseminated to the community for additional feedback. The project team will also track attendance for conference presentations to gauge interest in project activities and goals and to determine the characteristics of conference presentation attendees to define further CuRe Training Program audiences and their characteristics.

Project Resources

The project team requests $49,820 to cover costs for salary and fringe benefits, and travel support for team members to attend professional conferences to deliver reports and solicit feedback on project activities. The budgeted amount is inclusive of the UNC Facilities and Administration Cost Rate of 36.00% for on-campus sponsored activities other than organized research for instruction applied to direct costs. For full budget information, please refer to the Budget Form and Budget Justification documents.

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 Thu-Mai Christian, Assistant Director for Archives at the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (UNC), who will oversee the administration of the project. 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. 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.

Limor Peer, Associate Director for Research at the Institution for Social and Policy Studies (ISPS) at Yale University, will serve as a project consultant who will lend her expertise in data curation standards and workflows to help develop the curricular framework for the Data CuRe Training Program. 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, Green, & Stephenson, 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.

Florio Arguillas is a Research Associate at the Cornell Institute for Social and Economic Research (CISER) at Cornell University. In this capacity, Florio assists campus researchers in qualitative and quantitative data management and processing, assessing and mitigating disclosure risks, and use of statistical software packages. Florio also 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 will serve as a project consultant 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.

Project Timeline

The project design is based on a one-year award period of performance starting October 1, 2017 and ending September 30, 2018. Project activities will begin with the environmental scan, followed by curricular framework development, then strategic planning. Engagement with the data curation and research communities will take place throughout the majority of 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, relevant listservs, social media, and professional conferences. The project will also increase awareness of Data CuRe Training Program plans by soliciting community 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.

3. DIVERSITY PLAN

The University of North Carolina at Chapel Hill and the Odum Institute are committed to diversity as a core value by supporting intellectual freedom, creating and sustaining inclusive environments, and promoting civil and cordial discourse. Likewise, diversity is reflected in the multicultural perspectives of project team members, who are sympathetic to disparities in educational and professional attainment by underrepresented groups. Therefore, the project team is committed to addressing diversity in the following ways:

- Involve 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.
- Consider 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.
- Attract 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.
- Design an accessible curricular framework to allow for individuals with disabilities to participate fully in all components of training and education programs.

4. NATIONAL IMPACT

The Data CuRe Training Program planning grant activities will produce several important outcomes that will benefit the greater library and archives fields working to support the increasingly complex needs of the research communities they serve. By producing an evidence-based model of data curation for reproducibility education and a strategic plan for implementing the model, the proposed project will have the following impact:

- **Enable the development of training and education programs that are responsive to the demonstrated needs of academic librarians and data archivists, and the research communities they serve.** Plans for the development of the Data CuRe Training Program provide the blueprint for building workforce capacity for data curation for reproducibility

services. Because these plans are based on information collected from existing education programs and data quality framework workflows, academic libraries and data archives will be able to provide services that appropriately address users' needs.

- **Support feasible implementation of data curation for reproducibility training and education programs that offer well-defined lesson plans, relevant content, and effective delivery methods.** The Data CuRe Training Program curricular framework will include learning objectives that align with data curation for reproducibility practices along with samples of learning activities, lists of resources, and specifications for teaching equipment and tools. A strategic plan for implementing the curriculum will give academic libraries and data archives practical approaches for developing and implementing a successful Data CuRe Training Program at their institutions and demonstrate a potential for collaboration among librarians, archivists, data specialists, and technologists in support of the core scientific principle of reproducibility.

- **Expand academic library and data archive participation in the development and application of data curation standards and best practices that fully support reproducible research.** Engaging the professional community in planning grant activities will help increase awareness of the role of the academic library and data archive in the reproducible research landscape. By soliciting their feedback on project outputs, the eventual Data CuRe Training Program will be grounded in library and archive community consensus on data curation for reproducibility standards and best practices.

The Data CuRe Training Program project planning grant will support the first steps toward greater goals—and greater national impact that a full-fledged training program can offer. A workforce of librarians and archivists prepared to work alongside researchers in their attempts to meet rigorous standards for data quality will directly impact the reproducibility of scientific research. Moreover, it allows libraries and archives to continue to uphold their mission to provide access to the collections, expertise, tools, and services that support scholarship.

Laura Bush 21st Century Librarian Program (LB21-FY17-2)
Odum Institute, University of North Carolina at Chapel Hill

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 October 1, 2017 and ending September 30, 2018.

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
1. Environmental scan												
Analyze existing data curation for reproducibility workflows												
Examine current data curation education and training programs												
Review computational statistics training programs												
2. Curricular framework development												
Develop Data CuRe curricular framework												
Solicit feedback on curricular framework from CURE Consortium advisory group and data curation community												
Revise Data CuRe curricular framework based on feedback												
3. Strategic planning												
Develop strategic plan for curricular framework implementation												
Solicit feedback on strategic plan from CURE Consortium advisory group and data curation community												
Revise strategic plan on feedback												
4. Community engagement												
Expand CURE Consortium membership												
Disseminate project findings and outputs at professional conferences												
Disseminate project outputs on project website and data repository												

DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., 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. However, applying these principles to the development and management of digital products can be challenging. 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
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 format 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.

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)

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will 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 archived in the UNC Dataverse repository for long-term preservation and access during and after the award period of performance.

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/

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.

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