Building the Digital Curation Workforce: Advancing Specialized Data Curation

Abstract

Washington University in St. Louis (WU), in partnership with Cornell University, Duke University, Johns Hopkins University, the University of Michigan, and the University of Minnesota (members of the "Data Curation Network"), propose a two-year project (July 1, 2018 - June 30, 2020) to develop an intermediate to advanced level data curation training program open to all academic library staff. The project team is requesting $249,827 in support from IMLS. This program seeks to build and extend capacity for advanced data curation among academic library staff nationwide.

The past few years have seen an increased demand for subject liaisons, library archivists, digital curators, and library repository managers to incorporate data management and curation activities into the suite of services provided to faculty and students. Subject liaisons possess specialized discipline knowledge and expertise that can be leveraged to support the management and curation of research data. Yet, in practice, many lack the confidence and training to appropriately support researchers' demand. Additionally, repository managers, archivists and others tasked with curating research data in academic libraries are required to curate across many different disciplines and data types, often without extensive knowledge or expertise in all of those areas. This program will address both the need for front line library staff to gain confidence in research data-related services, and upskill the discipline knowledge of research data, repository, and archival staff, through a peer-to-peer training model.

The Specialized Data Curation Workshop program will:

1. Create intermediate to advanced data curation training modules and functional primers which will be made widely available for use and reuse.
2. Enhance capacity for librarians nationwide to robustly curate the heterogeneous datasets created on their campuses.
3. Expand the network of data curators, subjects specialists and repository managers nationwide.

Building the Digital Curation Workforce: Advancing Specialized Data Curation

Washington University in St. Louis (WU), in partnership with Cornell University, Duke University, Johns Hopkins University, the University of Michigan, and the University of Minnesota (members of the "Data Curation Network" or "DCN"), propose a two-year project (July 1, 2018 - June 30, 2020) to develop an intermediate to advanced level data curation training program open to all academic library staff. The project team is requesting $249,827 in support from IMLS. This program seeks to build and extend capacity for advanced data curation among academic library staff nationwide.

1. Statement of Broad Need

The past few years have seen an increased demand for subject liaisons, library archivists, digital curators, and library repository managers to incorporate data management and curation activities into the suite of services provided to faculty and students. Subject liaisons possess specialized discipline knowledge and expertise that can be leveraged to support the management and curation of research data. Yet, in practice, many lack the training, and therefore confidence, to appropriately support researchers' demand. Additionally, repository managers, archivists and others tasked with curating research data in academic libraries are required to curate across many different disciplines and data types, often without extensive knowledge or expertise in all of those areas. This program will address both the need for front line library staff to gain confidence in research data-related services, and upskill the discipline knowledge of research data, repository, and archival staff, through a peer-to-peer training model.

Results from our recent SPEC Kit #354 (Hudson-Vitale, et. al., 2017) indicate that library staffing for data curation is one of the most significant challenges facing academic libraries. Many institutions rely on the equivalent of less than one full time staff member to provide research data management and curation support. Despite this lack of resources, data curation services have been repeatedly identified in reports and publications as a rapidly growing and critical service area for the transformation of academic libraries. Additionally, research indicates that faculty and researchers require a variety of curation treatments for their data, but often lack the support needed to satisfactorily complete the treatments (Johnston et. al., in press).

Librarians and archivists understand the value and challenges of creating and preserving information for future generations, and recognize that specialized curatorial actions must be taken to preserve data and other materials in our institutional and disciplinary repositories for reuse. This curation enables data discovery and retrieval, maintains data quality, adds value, and provides for reuse over time through activities including authentication, archiving, management, preservation, and representation. Thus data curation is central to our mission and has become an important role for academic research libraries as we transform our workforce to assume greater digital stewardship responsibilities in the academy (Hedstrom, 2015). However our research with controlled pilots of data curation for the same dataset across six academic institutions uncovered inconsistencies in the application of data curation activities (DCN, 2017b). A subsequent follow-up survey of 80 academic libraries connected these inconsistencies to a deficit of well-trained curation staff (Hudson-Vitale et al., 2017). To address these inconsistencies, the DCN will also provide normalized curation practices and professional development training for the larger data community.

Currently, there are several digital curation training opportunities available as continuing education for library and archive professionals. Educational preparation for data curation services, like the DigCCuRR Professional Institute (https://ils.unc.edu/digccurr/institute.html) and the CLIR data curation post-doctoral fellowship program (https://www.clir.org/fellowships/postdoc), as well as information sharing networks such as the Digital Liberal Arts Exchange (https://dlaexchange.wordpress.com) and the DataQ Project (http://researchdataq.org) provide models for data curator workforce training. The Digital Archives Specialist (DAS) program developed by the Society for American Archivists focuses on the tools and technology needed to preserve digital assets, while the Digital POWRR program centers on foundational digital preservation training for staff from small to mid-sized institutions. Additionally, the recently funded IMLS planning grant "Data Curation for Reproducibility Training Program" is focused on developing training around datasets to ensure computational reproducibility.

According to a recent CLIR report (Allard et al., 2016), these existing curriculum materials "could be adapted, extended, or built upon to expand and scale up current educational and training offerings to address increasing needs." Specifically, our program addresses the CLIR report recommendation to collaboratively develop advanced and specialized interdisciplinary, competency-building curriculum and training. The National Research Council 2015 report, *Preparing the Workforce for Digital Curation*, provides an analysis of the curricula learning outcomes needed for training a digital curator workforce. Our work will build on this by developing intermediate to advanced data curator skills. Our co-PI instructors bring expertise in this area as directors of institutional data repository and curation programs across six major universities, as well as leads on past projects such as the Data Curation Profiles ToolKit and the Data Information Literacy (DIL) project that developed 12 competencies for teaching STEM graduate students data management skills.

This project will harness the experience of the the Data Curation Network. The DCN project was launched with funding from the Alfred P. Sloan Foundation in 2016 and seeks to develop a shared staffing model across nine institutions and organizations for data curation. By sharing our data curation expertise across a 'network of expertise' we aim to enable our academic libraries to collectively, and more effectively, curate a wider variety of data types (e.g., discipline, file format, etc.) that expands beyond what any single institution might offer alone. Training and community-building is a natural component of our vision and, therefore, the next logical step is to expand this training to a wider audience of library and archive practitioners. The expertise across the Data Curation Network partners will enable this project to provide in-depth training and reach a wide audience of librarian and archivist staff at national research libraries. As a result of sharing tools, providing a community of practice for data curators, and promoting data curation best practices, the DCN project seeks to establish a national platform for continuing education in digital data curation. It will lay the foundation to build an innovative community that enriches capacities for data curation writ large across the profession.

Research by the DCN demonstrated that both subject and functional expertise are needed to richly curate data that are being deposited in our institutional or data repositories (DCN, 2017c). During the DCN planning phase (May 2016-June 2017) six institutions observed 24 different files types from 176 datasets (see Table 1). These files types also spanned over 76 different disciplines, the most frequent being Anthropology (n=15), Crop Sciences (n=12), Civil, Environmental and Geo-Engineering (n=9), and Oceanography (n=7).

Table 1: DCN Reported file types sorted by most to least frequent (n=176)

Primary File Type	Count
Tabular	51
Sequence	25
Code Matlab	5
Video	5
Database SQL	1
Database	1

Geospatial	19
Image 3D	16
Code R	9
Code Python	7
Code Java	6
Audio	6
Image Biological	5

Textual	4
Survey SAS	3
Simulation	3
Database Access	3
Survey SPSS	2
Spectra	2
JSON	2

Plots	1
Autocad	1
Image	1
Code C++	1

While tabular data has widespread use, its curation is made more difficult depending upon the subject matter. For example, one of the tabular datasets received during this period was comprised of instrument readings of voltage-clamp fluorometry of heart valves. While the files are easy to open and view, without some type of discipline knowledge it is difficult to know if they are complete and well described, which complicates the curation. On the flip side, one of the datasets reflected in the numbers above is comprised of metadata about international roll call votes in a SQL database. While this type of metadata is relatively easy to understand, without knowledge of how to curate or understand a SQL file, they are inaccessible for curation. While extensive training and experience in SQL or voltage-clamp fluorometry is not required to richly curate research data, a baseline knowledge of how to open, check, and inspect the files is extremely helpful.

Ultimately, the training provided by this program and the capstone outcomes will extend the expertise of library staff into data curation, while building capacity for the curation of discipline-specific data among data curators and research data management staff found across the library landscape.

2. Project Design

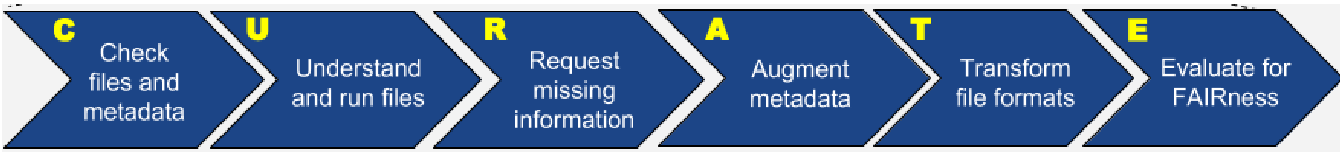
Our approach will leverage the in-depth digital data curation training for data curators established for the DCN (with hands-on curation procedures for a variety of disciplinary subject areas). The project grant is unique as it will build on successful digital curator training competencies developed in past IMLS projects (CLIR pub. 174) but innovate these past approaches by deepening the focus in the following ways:

1. Focus on intermediate to advanced skills for data curation from a discipline and functional perspective
2. Prepare participants to work with the researcher much earlier in the research process
3. Provide mock training of what could happen in the faculty's office; build confidence among curators and library staff to talk about data issues.
4. Build from Data Information Literacy competencies and other researcher/discipline curricula; builds on archival theory and best practice and applies them to data.
5. Incorporate the experiences and expertise of participants to enrich the curriculum and iterate the training content each subsequent session that is offered.

A preliminary version of the Advanced Data Curation Workshop curriculum was piloted in the winter of 2017 with funding from IASSIST and DCN member institutions. This 1.5 day workshop brought together approximately 45 librarians to learn how to richly curate research data from a number of domains. These professionals came from a wide variety of functional backgrounds, including catalogers, archivists, subject specialists, repository managers, and more. Attendees traveled to St. Louis, MO from as far away as the University of Virginia in Charlottesville, VA and Boise State University in Boise in Idaho. Additionally, there was a diverse makeup of institutions represented, including liberal art colleges, R1 institutions, and government agencies.

This pilot workshop focused heavily on the hands-on treatments required to richly curate research data. Attendees learned and then practiced a six step CURATE process of checking, understanding, requesting, augmenting, transforming, and evaluating datasets to be Findable, Accessible, Interoperable, and Reusable (FAIR) (Wilkinson, et al., 2016). Additional lessons focused on the value of data curation and introducing attendees to IASSIST and the DCN. Each attendee left the workshop with experience in curating real-world data, an elevator speech for talking about curation needs, and a plan for moving curation services forward or enhancing existing services locally. An outline of the workshop curriculum is available in the Appendix.

Figure 1: CURATE method for curating datasets



Feedback from the pilot indicated that the content was appropriate for the experience level of attendees. While we focused heavily on the manual treatments required to richly curate data, many attendees wanted more exposure and experience with software and tools they could use to curate data. Additionally, the hands-on portion of the workshop was extremely successful given the use of real-world datasets from each of our institutions. Using actual submissions highlighted the complexities of curating data an institution may not immediately have the expertise to curate and ultimately helped to shape the development of this program's capstone output.

Extending the earlier pilot curriculum, we will focus on peer-to-peer training by staff tasked with data curation to share and expand their expertise in a variety of data curation techniques for multiple data file types (e.g., geospatial, spreadsheet/tabular, statistical/survey, audio/video, computer code) and discipline-specific datasets (e.g., genomic sequence, chemical spectra, historical papers, biological images) in order to enable academic institutions to better support researchers that are faced with a growing number of requirements to ethically share their research data. This will be undertaken through the capstone project whose output will be a set of data curation primers. For this capstone, data curators and subject experts will form small groups of 3-4 to create documents that detail how certain data formats and specialized datasets could be curated. This should build upon experience and expertise the group of librarians already possess. For example, a group might address what considerations need to be taken into account in order to curate .czi files - which are a common, proprietary file format outputted by microscopes in a variety of disciplines. These considerations would then coalesce into a document that other curators or librarians would use when they receive a .czi file at their institution and are preparing the files for curation.

Our intended audience are librarian and archivist staff involved in digital data curation efforts at our nation's academic research libraries. These staff manage digital repositories, hold subject-specific duties related to data curation, consult with researchers on data management plans, and curate a wide variety of digital objects. Yet, many may find that curating digital data poses a greater challenge than less complex digital objects or that new data sets present new challenges every day (Hudson-Vitale et al., 2017). By bringing together curators from a wide variety of disciplines and specialty focuses, our project will allow these professionals to share curation techniques at the disciplinary and file-type level and build a community of curators that can thrive post-event.

Projected Performance Goals and Outcomes: The project will: (1) expand functional data curation capabilities for librarians; (2) enhance the quality of data curated at institutions that participate in these programs, and (3) create discipline-specific or functional primers. The following performance goals will measure the progress and success in achieving these outcomes: (1) capstone projects and activities from participants; and, (2) reuse of primers indicated through downloads and citation counts.

Recruitment and Selection: The Advanced Data Curation Workshops will be open to all librarians with a basic level of data management or curation experience. A call will be placed on public, library-focused listservs, including Code4Lib, the ACRL digital curation, and DataCure, among others. Each attendee will be asked to complete a brief application in which they will indicate their background and experience in providing or studying research data management or curation. Special attention will be given and explicit efforts will be made to recruit participants from underrepresented groups and institutions with limited resources. In an effort to expand the diversity of attendees, 10 scholarships will be made available to help support individuals who could not otherwise afford to attend.

Workshop Format & Content: This program will consist of a 2 day training event for 25 participants every 6 months over the course of the two-year project - for a total of three workshops. The training will focus on in-depth *functional* and *discipline-based* data curation skills and competencies (rather than on the administrative or managerial skills needed to develop and execute a robust digital curation program).

Specifically, the project team will draw upon the results of our previous research (Johnston et al., 2018), in which faculty and researchers identified the most important curation activities for their work and the treatments for which they were least satisfied. This research indicated a number of high priority areas for data curation treatments as indicated in Table 2. A draft of the curriculum is included in the "Curriculum" supporting documentation.

Table 2: Researcher data curation activities

Data Curation Treatment	**Definitions**
Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).
Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.
Quality Assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the

	significance of “null” and “blank” values, or unclear acronyms.
Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.
Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).
Technology Monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.
File Audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.
Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).
Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.
Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.
File Format Transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.

Capstone Project: At the conclusion of the workshop, participants will be formed into small groups of 3-4 individuals and asked to draft a capstone project plan. This capstone project will be due 6 months later, in the form of a discipline-specific or functional primer to be developed into a open educational resource. There will be flexibility in what each of these primers look like - for example it could be a curator checklist for dealing with a specific file format or a tool to help curators evaluate a data type appropriately. One required component of the capstone will be for participants to interview 1-2 researchers with a similar data type or whose data they are curating about their research workflows, using a modified version of the Data Curation Profile (Witt et al., 2009) to ensure the modules are based on real-world processes and data. We will iterate this approach three times, thus building a more comprehensive training program garnered from the combined skills and expertise of the practitioner participants who attend each session. The project will output a final set of discipline and functional primers and training modules that will be shared with the community at large as open resources. At a minimum we see these primers each containing the following information about their topic:

- File format background and applications
- Tools for opening file type (proprietary and open source)
- Transformation considerations
- Curation workflows/steps
- Recommended metadata schema and elements
- Preservation recommendations
- Links to example curated dataset

Cohort Development

Building a cohesive community of intermediate to advanced data curators is of significant importance given the specialization often required to robustly curate the heterogeneous data found on university campuses nationwide. To enhance this community, the training program will borrow from the 2016, Association of Research Libraries and Center for Open Science SHARE Curation Associates pilot program which supported the curation of digital assets through a community development and training program for mid-career library staff. The SHARE pilot program successfully brought together over thirty geographically dispersed library professionals to enhance local digital assets and learn computational techniques, while building community around the curation of metadata.

Specifically, the Advanced Data Curation Workshops will put in place a number of communication channels and virtual collaboration spaces to facilitate community.

- *Monthly Small-Group Calls:* After each workshop attendees will participate in monthly calls for the next 6 months. These calls will help attendees keep on track to complete their primers within 6 months and ensure the successful deployment of the primers after their completion.
- *Open Science Framework (OSF) Project Space:* OSF is a free, secure web application for project management, collaboration, registration, and archiving. The workshop attendees will have a dedicated space on the OSF to track projects and activities, access meeting notes, review training materials, and communicate about curation and computational efforts.

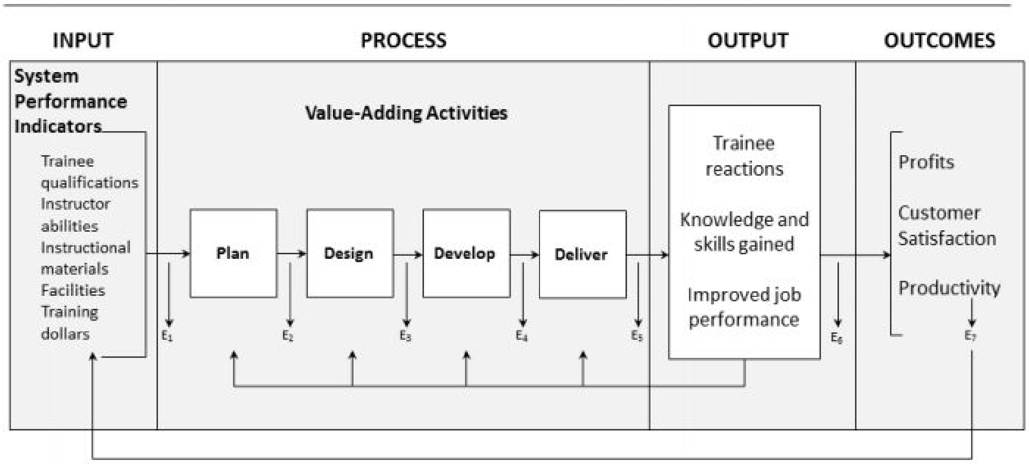
Timeline

August 2018	Curriculum and instructor meeting; Location: Washington University in St. Louis, St. Louis, MO
October 2018	Advanced Data Curation Workshop #1; Location: Digital Library Federation Forum, Las Vegas, NV
April 2019	Advanced Data Curation Workshop #2; Location: Johns Hopkins University, Baltimore, MD 1st cohort capstone project due.
October 2019	Advanced Data Curation Workshop #3; Location: Washington University in St. Louis, St. Louis, MO 2nd cohort capstone project due.
April 2020	Final curriculum development meeting; Location: Washington University in St. Louis, St. Louis, MO 3rd cohort capstone project due.
June 2020	Release full curriculum and final report

Assessment

To assess the effectiveness of this workshop, the team will use a modified systems-based Input, Process, Output, Outcome (IPO) Model (Bushnell, 1990) as an evaluation framework (Figure 2). Each component of this model is defined as such: i) Input: system performance indicators such as "attendee qualifications and the availability of already tested instructional materials"; ii) Process: implementation of the training through establishing objectives and developing material; iii) Output: data resulting from the training iv) Outcomes: longer-term results associated with training. We will gather feedback from the internal team (E1-E4) and participants (E5-E7) in the training implementation. E1-E4 will be incorporated into the development process with the Team, while the other evaluations will occur through exercises (E6) and surveys (E5-E7). Training and evaluation questionnaires informed by Grohmann's motivation assessments (Grohmann, 2013) and Giangreco's, et al. (Giangreco, 2009) reaction assessment will be used as evaluation surveys. All shared modules/information will be evaluated by usage counted from site analytics.

Figure 2: Assessment Model



Project Resources and Personnel

Key Project Personnel: (See "Project Staff" and "Resumes" for more information.)

(1) **Cynthia Hudson-Vitale**, Data Services Coordinator at WU, will act as Project Director, overseeing project management and implementation. She will be responsible for hiring and managing the Project Manager, and coordinating support for the primers. Ms. Hudson-Vitale will help recruit attendees, develop curriculum, and provide primer development oversight to attendees. She will also manage IMLS reporting requirements and the dissemination of information about the fellowship.

(2) **Project Manager, To be hired:** (.5 FTE for 2 years: $54,800 in grant funds). WU will hire a part-time Project Manager to start in June 2018. This position will handle the day-to-day management of the workshop program, including scheduling training and monthly meetings, working with attendees on their primers, and the development of open educational resources to be released after the workshop concludes. Additionally, this individual will explore sustainability options for continuing the program at the end of the grant period.

(3) **Mara Blake,** is Data Services Manager at Johns Hopkins University. Ms. Blake will help recruit attendees, develop curriculum, and provide primer development oversight to attendees.

(4) **Jacob Carlson,** Director of Research Data Services, University of Michigan. Mr. Carlson will will help recruit attendees, develop curriculum, conduct the data curation workshop training, and provide primer development oversight to attendees. .

(5) **Joel Herndon, Ph.D.** is head of data and visualization services, at Duke University. Dr. Herndon will help recruit attendees, develop curriculum, and provide primer development oversight to attendees. .

(6) **Lisa Johnston** is the Research Data Management/Curation Lead at the University of Minnesota. Ms. Johnston will help recruit attendees, develop curriculum, conduct the data curation workshop training, and provide primer development oversight to attendees.

(7) **Wendy Kozlowski** is the data curation librarian at Cornell University. Ms. Kozlowksi will help recruit attendees, develop curriculum, conduct the data curation workshop training, and provide primer development oversight to attendees.

Instructors: Key project staff or data curators from collaborating institutions will act as instructors for the workshops.

Facilities, Equipment, Materials, and Supplies: The Advanced Data Curation Workshop program requests $50,600 for food and beverage and for room rental for the workshops. See “Budget Justification” for a breakdown of these costs.

3. Diversity Plan

Early career, non-traditional librarians and curators will be encouraged to attend. The project team is committed to maximizing the recruitment and retention of women, underrepresented racial and ethnic groups, and people with disabilities. All of the partnering institutions have ongoing programs that focus on the recruitment and retention of trainees and faculty from diverse ethnic backgrounds that are typically underrepresented. These programs also target individuals from disadvantaged backgrounds and people with disabilities. Enhancing diversity at WU is the top strategic goal approved by the Board of Trustees and a primary mandate of the Office of the Provost. A recent initiative includes the creation of the WU’s Center for Diversity and Inclusion to support students from traditionally underrepresented or marginalized populations and to promote dialogue through cross-campus partnerships. Specifically, we will:

- offer 10 scholarships of up to $1500 each to support the attendance of minority, disadvantaged, and underrepresented individuals;
- develop educational modules and primers that are fully accessible to individuals with disabilities;
- disseminate the educational modules and primers widely to extend their reach to a global audience;
- include a diverse array of data types and disciplines in the primer development.

4. Broad Impact

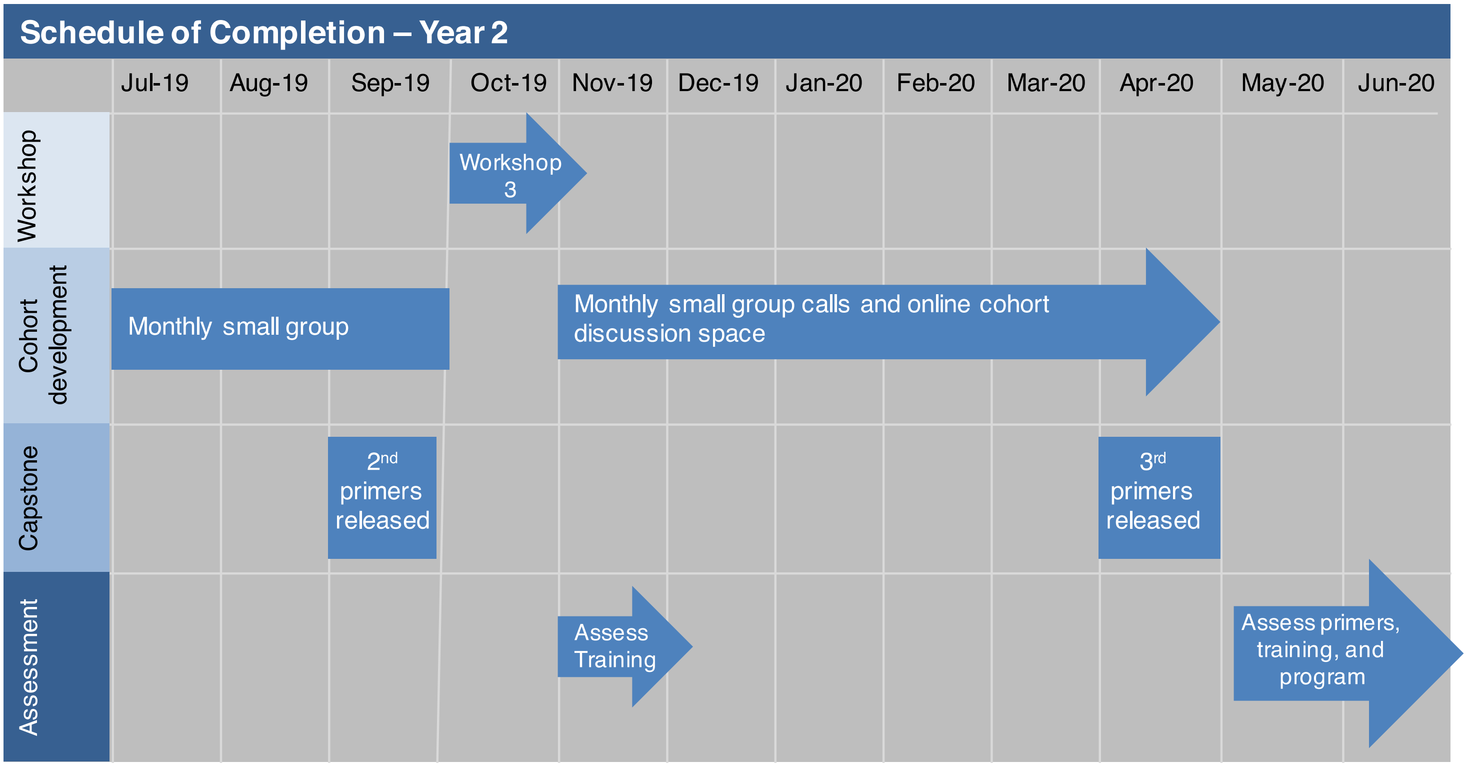
The Advanced Data Curation Workshop program will:

1. Create intermediate to advanced data curation training modules and functional primers which will be made widely available for use and reuse.
2. Enhance capacity for librarians nationwide to robustly curate the heterogeneous datasets created on their campuses.
3. Expand the network of data curators, subjects specialists and repository managers nationwide.

Schedule of Completion – Year 1

	Jul-18	Aug-18	Sep-18	Oct-18	Nov-18	Dec-18	Jan-19	Feb-19	Mar-19	Apr-19	May-19	Jun-19
Curriculum Dev & Refine		Meeting 1			Curriculum refinement						Curriculum refinement	
Workshop				Workshop 1						Workshop 2		
Cohort Development					Monthly small group calls and online cohort discussion space						Monthly small group calls	
Capstone									1st primers released			
Assessment				Assess Training						Assess Training		

Schedule of Completion – Year 2



	Jul-19	Aug-19	Sep-19	Oct-19	Nov-19	Dec-19	Jan-20	Feb-20	Mar-20	Apr-20	May-20	Jun-20
Workshop				Workshop 3								
Cohort development	Monthly small group				Monthly small group calls and online cohort discussion space							
Capstone			2nd primers released							3rd primers released		
Assessment					Assess Training						Assess primers, training, and program	

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

□ 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.

There will be multiple intellectual outputs of this training program. They include: (1) capstone projects; (2) training materials; and, (3) results from assessments. All assets will be released with a Creative Commons license for data sets or reports and GNU or MIT licenses for software of code. These licenses were chosen based upon their simplicity and dedication to openness and contribution to community efforts. Assessment results will be de-identified and made available through the OSF project space in the public domain. All curriculum developed through this program will be made available for reuse through a Creative Commons license.

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.

Digital products will be released to the public domain for wide distribution and reuse.

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 only assets that will include privacy concerns is the results of the in-depth assessments and surveys. These materials will be de-identified and aggregated at a level where re-identification will be impossible.

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.

Digital content produced through this program may include monthly meeting notes and primer documents. Most of these assets will be in the form of text documents, but may include interactive websites, powerpoint presentations, and more.

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.
Standard text editors will be used to create the textual documents, such as Microsoft Word, Texteditor, or other open source, non-proprietary text creator.

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).
TXT or .docx files will be used for producing these assets.

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).
Six months after each cycle of workshop the discipline or functional primer will be released for reuse and public feedback. Through the OSF, versioning of assets is automatically documented, in addition to the creation of unique persistence identifiers.

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 COS has a robust preservation plan for assets stored on the OSF. Through the OSF, the assets will be aggregated into the SHARE data set where they may be linked, curated and enhanced through machine learning techniques and human intervention. In addition to txt or .docx files, the assets will be transformed into PDF/A documents to aid in accessibility. Copies of the training materials and primers will also be stored on the WU institutional repository. Thus providing redundant storage and preservation of the asset.

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).

Descriptive metadata (Dublin Core) will be created through both manual and automatic means. Meeting notes will track provenance through versionning on the OSF. Each assignment will automatically assign a creator and upload date, and additonal fields for description, tags, and the ability to assign a license. Local practices for metadata will vary by institution, but at a minumum, we anticipate the creation of Dublin Core and PREMIS metadata records.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.
Metadata will be stored with the digital assets on the OSF and the local institutional repository. The dual storage of the assets both on the OSF and through the institutional repository ensures the ongoing availaibility of the metadata and asset. Checksums will be created for both the metadata records and the assets to protect against bitrot and any changes to the metadata files.

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).

All the metadata will be fed into the SHARE data set for greater discoverability. SHARE is aggregating metadata about scholarship, enhacing and curating the metadata, and finally making it available through a common API.

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).

The assets will be made widely available through the Open Science Framework and the WU institutional repository.

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.

Examples of meeting notes and assets created through the SHARE pilot program may be found here: http://osf.io/c3veb

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.

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.

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.

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

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

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.

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.

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

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

Name of publicly accessible source code repository:

URL:

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.

This program will create data sets as a result of institutional assessments and potentially through fellow capstone projects.

Assessment data will be collected through surveys. The surveys will take place after each workshop. Additionally, a survey will be administered at the end of the fellowship to receive overall program feedback. Exit interviews will also be conducted at the end of the fellowship.

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?

The assessment data may require IRB approval, upon grant funding, the project team will secure approval through the WU IRB.

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).

No PII, direct, or indirect identifiers will be collected through these data sets.

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.

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).

The survey data will be collected through a common survey software such as SurveyMonkey, Google Forms, or Qualtrics. Each of these tools have exporting features that allow the exports in common or non propietary formats (excel or csv), thus there are no technical requirements or dependencies for viewing.

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?

In addition to the results of the survey, the survey instrument will be made available. These documents will be stored in non-proprietary or common text file formats (such as txt or .docx). These materials will be linked through the OSF project page and related to one another in a zip file. While digital permanence is questionable for any digital asset, all efforts will be made to ensure the ongoing relationship between the dataset and the documentation. These efforts may include scheduling file fixity checks, using URI's to associate materials rather than text in field boxes, and backing up both asset and documentation.

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?
An SIP, AIP and DIP will be created for assessment data created through this program. Following the OAIS model, the SIP will include the original documentation and data set to be stored in the WU Data & GIS Services. The AIP will include a copy of the SIP, the DIP, and the curatorial outputs of the Bitcurator program, which includes PREMIS metadata and checksums. The DIP will be placed on the WU Open Scholarship repository and include access copies of the assets, any code, the interview instrument, and the Dublin Core metadata record. According to WU Open Scholarship data policies, after 10 years the data will undergo a collection review by the appropriate subject librarian.

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

Name of repository: WU Open Scholarship (for assessment data)

URL: openscholarship.wustl.edu/data

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

The data management plan will be reviewed every month, along with project goals and milestones. The implementation will be monitored by the Project Director and any changes will be evaluated by project team.