LG-72-17-0139-17 Trustees of Indiana University

Abstract

The research team from the School of Informatics and Computing (SoIC) at Indiana University Indianapolis (IUPUI) and Indiana University Bloomington (IUB) will develop a conceptual framework for assessing libraries' capacity for big data curation. The purpose of this planning project is to provide a research-based foundation for a full project, which involves the development of a big data curation capacity assessment toolkit that will be freely available to academic and public libraries. The conceptual framework will be informed by three specific research activities: (1) systematic literature review of institutional/organizational capacity and big data and data curation literatures, (2) development and administration of an online survey to understand current library data curation practices, and (3) in-person focus groups on the topic of big data curation capacity in academic and public libraries with an expert panel.

This planning project will have a major impact on libraries, both academic and public. A conceptual framework will be essential in implementing sustainable and scalable big data curation programs, and it will assist library staff (e.g., data curators, data librarians, and library administration) in understanding their current environments as well as potential impediments to building successful curation programs. By better understanding their capacity for big data curation, the outcomes of the project should have a direct impact on their work, in particular their efforts to implement effective big data curation programs across the country. As project outputs, the project team will generate a bibliography for capacity and big data curation; collect information from library professionals about their current curation practices in libraries; and create a conceptual framework for assessing libraries' capacity for big data curation while specifying the necessary conditions for increasing it. These outputs will provide informative sources for academic and public libraries that are interested in or are planning curation programs.

Library Capacity Assessment and Development for Big Data Curation

Project Overview

The research team from Indiana University Indianapolis (IUPUI) and Indiana University Bloomington (IUB) will develop a conceptual framework for assessing libraries' capacity for big data curation. This conceptual framework will be informed by three specific research activities: (1) systematic literature review of institutional/organizational capacity and big data and data curation literatures, (2) development and administration of an online survey to understand current library data curation practices, and (3) in-person focus groups on the topic of big data curation capacity in academic and public libraries with an expert panel. The purpose of this planning project is to provide a research-based foundation for a full project, which involves the development of a big data curation capacity assessment toolkit that will be freely available to academic and public libraries.

STATEMENT OF NEED

The Need for Big Data Curation in Libraries

Data are increasingly being recognized as "first-class intellectual objects" that support scientific inquiries and improve the quality of human life. Defined as "high-volume, high-velocity and/or high-variety information assets" (Gartner, n.d., n.p.), big data also emerges as the basis of modern research in many areas such as medicine, environmental science, and urban planning (Reinhalter & Wittmann, 2014). Reinhalter and Wittmann (2014) argued that this poises librarians to be a vital part of big data stewardship, including big data storage, processing, and use. Traditionally libraries have been responsible for curating data to address their mission by protecting and disseminating data (Heidorn, 2015). Now libraries need to play a role in curating and providing access to big data (Teets & Goldner, 2013). This is not just for academic and research libraries that provide data management services for faculty due to their production and use of big data, but also for public libraries. Given the rise in publicly available data resources, many open government data initiatives at federal, state, and city levels currently exist or are in the process of being created, and these initiatives involve partnerships with public libraries. For example, the Boston Public Library has a recent initiative to develop an open data collection for the City of Boston (City of Boston, 2017).

Because big data are fundamentally different from "spreadsheet science," curating them requires "innovative forms of processing that enable enhanced insight, decision-making and process automation" (Gartner, n.d., n.p.). Librarians need to embrace a role in making big data more useful, visible, and accessible by creating taxonomies, designing metadata schemes, and systematizing retrieval methods (Bieraugel, 2013). Big data curation can include storing faculty scholarly research and making it accessible as well as storing and mounting faculty raw research data for others to use. Demand from researchers for these types of big data curation services is increasing (Hudson-Vitale et al., 2017). This is important because researchers are a key class of stakeholders whom libraries are expected to support and serve. Many have asserted the necessity of building sustainable and scalable data curation programs in libraries that will be effective in continuing to support communities and scholarship that may be overwhelmed by big data (NISO, 2013). Consequently, a holistic approach to understanding libraries' capacity for big data curation is needed.

The Need to Assess Library Capacity

The concept of organizational capacity has received considerable attention during the past several decades. Methods for assessing organizational capacity have been developed, and some frameworks for building organizational capacity have been proposed (De Vita & Fleming, 2001; McKinsey & Company, 2001, 2015; Nu'Man et al., 2007). While several studies have proposed slightly different definitions of organizational capacity, common among them is the emphasis on an organization's ability to implement or perform internally or externally generated expectations and outcomes from their organizational field (Barman & MacIndoe, 2012; European Commission, 2005; Horton et al., 2003; Sharpe, 2006). This suggests that the conscious development

of organizational capacity is a critical part of enabling an organization to fulfill its mission. Previous researchers have emphasized that assessment is an essential part of the process of building organizational capacity, arguing that assessing capacity is a prerequisite for the interlinked decisions made in any organization, such as strategic and operational choices, ongoing policy dialogue, and further capacity development processes (European Commission, 2005).

As the concept of organizational capacity has been developed and applied primarily in the context of nonprofit organizations, it is applicable to public and academic libraries. However, despite its usefulness and significance, capacity building or assessment has not garnered much attention in the library field (Carrigan, 2015). The concept of capacity can be used to assess the conditions that are necessary for a library to perform curation activities and realize its potential for success.

The interest in and need for data services is increasing rapidly in libraries, and more libraries—mostly academic but also a few public libraries—have been trying to establish data services. Several recent studies have identified an increase in launching data services or extending existing services (e.g., Cox & Pinfield, 2014; Yoon & Schultz, 2017; Xia & Wang, 2014), which are and/or should be accompanied by proper data curation programs. These studies also demonstrated a significant level of variation in their services and programs, perhaps due to their differing organizational capacities (Yoon & Schultz, 2017). To successfully and sustainably launch and maintain data services in libraries, data curation programs must be tailored to a library's existing capacity. Eventually, libraries need to build or grow their capacity in appropriate ways to sustain or extend their data curation programs in the long term. Doing so will ensure their effectiveness and help libraries avoid unintended consequences. Particularly, due to the complex and distinctive nature of big data, curation programs require a thorough assessment of sufficient capacity in various dimensions, such as technology infrastructure, policies, values, skills, culture, and leadership.

Gaps in Library Assessment Efforts and the Significance of This Planning Project

Academic libraries have a long-standing history of launching initiatives to conduct either a formal or informal assessment; however, assessments have only become systematic processes since the late 1990s, when various assessment methods, such as surveys or library collection analyses, were employed (Hiller & Self, 2004). Major library organizations, such as the Association of College and Research Libraries (ACRL) and the Association of Research Libraries (ARL), have argued that assessment is important for libraries, given the increasing focus on institutional effectiveness by accreditors and others in the higher education community. In its report, "The Value of Academic Libraries," ACRL noted that "librarians are increasingly called upon to document and articulate the value of academic and research libraries and their contribution to institutional mission and goals" (2010, p. 6). The same argument has been made for public libraries because they also need to demonstrate the value of libraries to society and their return on public investment (Carolyn, 2014). Consequently, many assessment efforts have been developed to document outcomes or investigate performance. Kaufman and Watstein (2008) argued, "libraries cannot demonstrate institutional value to maximum effect until they define outcomes of institutional relevance and then measure the degree to which they attain them" (p. 227). Although academic libraries have led major assessment efforts, public libraries have also established standardized indicators for measuring their success (Closter, 2015), for example, the work of the Public Library Association (PLA) Performance Measures Task Force launched in 2013.

This project, which aims to develop a library capacity assessment method, contributes to increased service and program effectiveness assessment/measurement at libraries generally. With a focus on big data curation, **capacity assessment is a critical core instrument for planning, monitoring, and evaluating programs before a library defines its outcomes or even launches its curation programs.** As previously stated, capacity assessment is a fundamental prerequisite for the effective realization of expected outcomes. Thus, *it is timely and significant to provide a method to assess library capacity for big data curation because many libraries have recently launched or have been planning to launch new data programs.* The proposed capacity assessment

model from this project will be different from and a complement to other existing assessment efforts that are typically based on outcomes and performance, such as LibQUAL+®, which is a user survey that helps libraries assess and improve library services, change organizational culture, and market the library, or SPEC surveys, which gather information on current research library practices and policies from ARL member institutions to help them compare practices and improve performance. ARL just published Data Curation, SPEC Kit 354, in May 2017, which explores the infrastructure that ARL member institutions are using for the active and ongoing management of data (http://publications.arl.org/Data-Curation-SPEC-Kit-354/). SPEC Kits usually report on current trends that are of interest to library researchers and practitioners. SPEC Kit 354 underscores the current issues that librarians need to address regarding data curation. Our work will be a valuable complement to Data Curation SPEC Kit 354 for two main reasons. First, in contrast to SPEC Kit 354, our study is not limited to ARL institutions. Second, SPEC Kit 354 focuses on current staffing and infrastructure (policy and technical) at ARL institutions for data curation, the level of demand for data curation services, and challenges that institutions face in regard to providing data curation services. In contrast, our study focuses on academic and public libraries' ability to perform tasks and produce outputs, to understand and solve problems, and to make informed choices related to big data curation, all of which are cornerstones of the capacity concept. Although we acknowledge that there may be more or less formal capacity assessment tools developed locally for individual institutions, this project will be the first initiative to develop a standardized capacity assessment tool for academic and public libraries for data curation. As a first step in this planning project, we aim to build a conceptual framework for capacity.

PROJECT DESIGN

Project Goals

The goal of this planning project is to develop a conceptual framework for assessing libraries' capacity for big data curation, which will be essential in implementing sustainable and scalable big data curation programs. Our focus is on academic and public libraries, considering the recent growth in data services as well as the emerging needs and interests in launching new data programs.

In developing an assessment framework, the overarching aim of this project is to define capacity in the context of big data curation and to identify its various dimensions from the existing literature on the topic found in various fields (e.g., organizational studies, development studies, economics, and education) within the context of libraries and big data curation based on the examination of current library practices. This initial investigation will lead to a larger project aimed at developing an assessment toolkit that will be freely usable by academic and public libraries. Because we understand the differences between academic and public libraries, in terms of workforces, priorities, and funding, our goal is to use the data that we collect and analyze during this project to guide us in developing a single framework that is applicable to both types of libraries or two separate frameworks for big data curation capacity: one for academic libraries, and one for public libraries.

The Audiences of This Project

Data curators and data librarians at academic and public libraries are the primary audiences for our project. By better understanding the capacity for big data curation, the outcomes of our project should have a direct impact on their work, in particular their efforts to implement effective big data curation programs across the country. Another audience for this project is library administration. For example, data curators and data librarians can utilize the framework that we develop in conversations with library administration to garner their support in developing capacity for big data curation. This project will also benefit researchers, both producers and reusers of big data, because our project is focused on making sure libraries understand their capacity for making big data useable and for storing data, including valuable raw data. In turn, this information should help data curators and data librarians provide support to data producers and reusers.

Intended Outcomes and Project Outputs

This planning project will establish a theoretical understanding of the concept of capacity and identify the dimensions that constitute organizational capacity. As project outputs, the project team will generate a bibliography for capacity and big data curation. In addition, we will collect information about the current curation practices in libraries. We will then create a conceptual framework for assessing libraries' capacity for big data curation while specifying the necessary conditions for increasing it. These outputs will provide informative sources for academic and public libraries that are interested in or are planning curation programs. See the list below for the detailed items:

(a) A theoretical definition of library capacity in the context of big data curation
(b) A list of dimensions or attributes that comprise library capacity
(c) A bibliography of the relevant literature in the area of capacity and big data curation
(d) A report from the landscape survey on current library curation practices
(e) A conceptual framework for assessing libraries' capacity for big data curation
(f) A list of necessary conditions to increase capacity

Project Activities & Methods

The major project activities include (1) conducting a systematic review of the literature on organizational capacity, data curation, and big data practices; (2) conducting a large scale landscape survey with libraries to learn about their current practices and/or future efforts planned; (3) developing a framework based on the findings of (1) and (2); and (4) vetting the framework with a panel of leading experts who handle various aspects of big data curation in academic and public libraries.

Table 1. Summary of Project Activities and Timeline with Lead Personnel

Time	Project Activities	Methods	Outputs*	Lead Personnel
Months 1–3	Literature review of organizational capacity	Systematic literature review	(a), (b), (c)	Yoon
Months 1–3	Literature review of big data curation practices	Systematic literature review	(a), (b), (c)	Donaldson
Months 4–9	Landscape survey	Online survey	(d)	Yoon
Month 9	Framework development		(e), (f)	Yoon & Donaldson
Month 10	Expert review	Focus groups	(e)	Donaldson
Months 11–12	Final report & full project planning			Yoon

*See the list of outputs with details on page 4.

1. Systematic Literature Review (Months 1–3)

Schedule:

- **Framing questions and identifying relevant literature (Month 1)**
- **Literature assessment & analysis (Months 1–2)**
- **Interpretation (Month 3)**

The project team will first conduct a systematic literature review in two areas: organizational capacity and big data/data curation. The purposes of this literature review are (1) to understand the meaning of capacity and

propose a definition contextualized in big data and data curation and (2) to identify high-level capacity attributes and dimensions that are applicable within a library data curation context. Systematic review is a good approach for our project because it is known as a structured method that can identify and critically evaluate/analyze a vast amount of research outcomes in a research field. A systematic review establishes to what extent the existing research has progressed to clarify a particular problem and is useful for identifying gaps, relations, and any contradictions in existing studies. Moreover, a systematic review is useful for proposing a new conceptualization or theory, in particular, for formulating general statements or an overarching conceptualization to comment on, extend, or develop a new theory (Baumeister & Leary, 1997; Bem, 1995; Cooper, 2003).

In this project, we will adopt and slightly modify the five steps of conducting a systematic review, originally proposed by Khan, Kunz, Kleijnen, and Antes (2003):

Step 1. Framing questions for a review;
Step 2. Identifying relevant work;
Step 3. Assessing the quality of studies;
Step 4. Summarizing the findings; and
Step 5. Interpreting the findings.

In Step 1, we will frame the questions based on our topics: **Topic 1**. The concept of institutional/organizational capacity and dimensions attributed/identified in previous studies and **Topic 2**. Big data and data curation.
In Step 2, to identify existing studies on organizational capacity, we will conduct a keyword and title search in major databases, such as ProQuest, EBSCO, and ERIC. This will help us to identify studies conducted in various disciplines, outside of organizational studies and development studies, which are two well-known fields for conducting studies of organizational capacity. In addition to journal searches, we will also conduct a Google search to find any reports prepared by international nonprofit organizations, such as the United Nations, and major consulting firms, such as McKinsey. We will use bibliography management software for managing all resources (e.g., Zotero). For Topic 2, we will examine major information and library science journals as well as information technology journals (e.g., the *Journal of the Association of Information Science and Technology*, the *International Journal of Digital Curation*, *Library and Information Science Research*, *Library Hi-Tech*, and the *International Federation of Library Associations and Institutions Journal*). We will also examine conference proceedings from major conferences related to data curation and preservation, such as the International Conference on Digital Preservation (*iPRES*) and the Society for Imaging Science and Technology Archiving Conference (e.g., *Archiving 2017*). Once resources are identified, we will start collecting key information from the literatures (e.g., definitions, dimensions, and context) and analyzing the literature for its significance and impact while also weeding literature that is not relevant or of good quality (Step 3). The outcomes of this assessment and analysis will be summarized, and we will generate our interpretation of the concepts and dimensions of capacity to provide a foundation for our future work (Steps 4 & 5).

2. Landscape Survey (Months 4–9)
Schedule:

- **Survey design, review, and IRB (Months 4–5)**
- **Data collection (Months 6–7)**
- **Data analysis (Months 8–9)**

During this time, we will develop an online survey to understand current library data curation practices. This survey will complement our understanding of the relevant literature. The purposes of the survey are: (1) to identify libraries that already have big data curation programs or who are interested in developing such programs; (2) to understand the current curation workflow at academic and public libraries; (3) to identify human, technical, and other resources that are currently utilized for big data curation; (4) to investigate the challenges of current curation practices; and (5) to map current practices with the capacity dimensions identified from the literature review in the next step of the project (See 3. Framework Development and Vetting

Framework with Experts). The survey items will include Likert-scale, categorical, and open-ended questions. To ensure validity, the survey questionnaires will be reviewed by our advisory board.

Our sample will be drawn from the *American Library Directory: 2016–2017* (American Library Directory, 2016), which lists 16,878 public libraries and 3,635 academic libraries. We will randomly select libraries from these lists using the contact information provided in the directory to recruit contacts who are responsible for data curation at those institutions as well as those who in the beginning stages of considering a big data curation program and those who have already cancelled their data programs. Since academic libraries and public libraries have vastly different workforces, priorities, and funding streams, we will employ stratified sampling for our study. We will create strata for public libraries according to categories specified in the *American Library Directory*: (1) public libraries, excluding branches (n=9,669), (2) main public libraries that have branches (n=1,409), and (3) public library branches (n=7,209). Likewise, we will create strata for academic libraries according to categories specified in the *American Library Directory*: (1) community college libraries (n=1,115), and (2) university and college libraries (n=2,520). After sampling from each stratum, we will utilize weighted averages if necessary to ensure that no particular subset of academic or public libraries is overrepresented. We will distribute the online survey using SurveyMonkey under an IUPUI SoIC account. In the invitation email, we will ask the dean or director of each library to forward the invitation to their librarians who are involved with or have at some point considered any curation activities. In addition to the personal invitation to the survey, we will also distribute the survey through the relevant library association LISTSERVs, such as Research Data Access and Preservation Submit (RDAP), the PLA, the American Library Association (ALA), and ACRL. We will state the significance of the project in the survey invitation and will also report back the survey results to participating libraries. In an effort to produce generalizable results, we expect to send out survey participation invitations to 50% of both academic and public library populations. Based on recent survey response rate research, we anticipate an approximate 10% response rate (Dillman et al., 2014).

The collected data will be analyzed quantitatively and qualitatively. The data will be exported in CSV format and imported to SPSS or STATA for descriptive analysis, using univariate and bivariate descriptive statistics (e.g., frequency distribution). As this planning project is an exploratory study, a qualitative analysis of the responses to the open-ended questions will also be important and will be analyzed using MS Excel.

3. Framework Development and Vetting Framework with Experts (Month 9–10)
Schedule:

- **Initial draft of framework (Month 9)**
- **Expert review (Month 10)**

Outputs from 1 and 2 will enable us to define the term *capacity* as it pertains to big data curation by mapping the current practices with organizational capacity concepts and dimensions and will be used to produce a framework for institutional capacity for big data curation. To ensure that our framework is rigorously reviewed by a panel of experts, we will conduct a focus group with our advisory board. In month 10, our advisory board will travel to Indianapolis, Indiana, to meet for an in-person focus group. Prior to the meeting, they will receive pre-focus group materials (e.g., background, meeting schedule, and our framework). During the meeting, we will present them with our framework and ask them to examine it. We will also ask them to discuss its strengths and weaknesses. Finally, we will ask them to propose any changes to our framework as well as the potential of this framework to be applicable in different contexts (e.g., both public and academic libraries and beyond or not). Using the focus group method will provide us with valuable feedback by allowing us to observe the extent to which our advisory board members agree and have differences of opinion about our framework.

4. Final version of framework, Output Dissemination, and Full Project Planning (Months 11–12)
Schedule:

- **Final version of framework (Month 11)**

- **Final report writing (Months 11–12)**
- **Full project planning (Month 12)**

The project team will create a revised, final version of the framework by consolidating the findings from our literature review, surveys, and focus groups carried out during Months 1–10. As explained earlier, the framework will provide a definition of library capacity in the context of big data curation and identify the dimensions and attributes of capacity. We will either have one framework for both public and academic libraries or two separate frameworks for each, depending upon the results of our study and data analysis. The project team will also write a report to submit to IMLS and prepare to disseminate the project findings and outputs of the project (e.g., project information, our preliminary work, or earlier outputs) through multiple channels, although output dissemination will take place throughout the project periods. During the last month, we will start planning a full project to develop a toolkit for assessing library capacity for big data curation. Although the details of the full project will be guided by the outcomes/outputs of this planning project, we envision that the full project components will include three major parts: (1) a community evaluation on capacity criteria from library professionals and pilot testing; (2) a national scale capacity assessment with public and academic libraries, using our network of libraries built in this planning grant; and (3) recommendations for building capacity.

Evaluation Plan

The project team will use several strategies to evaluate this planning project and the outputs. Because this project includes research components, we will examine the validity and reliability of the data collected from our surveys. Our eight advisory board members will check whether the survey items are appropriate for achieving the research purposes and ensuring the validity of the data. Internal reliability will also be checked for survey responses. Our advisory board, which includes representative experts from the digital curation discipline from public and academic libraries, curation technology, and the field of organizational capacity (see Project Resources for our advisory board), will play a role throughout the project period by providing feedback on our project progress and outputs. For instance, after creating the conceptual framework, we will invite our advisory board members to come to Indianapolis to review our framework. The advisory board members will also participate in a focus group aimed at generating a consensus on what constitutes capacity for big curation while considering our framework (see Project Design: 3. Framework Development and Vetting Framework with Experts). After a focus group meeting, the advisory board and the project team will have an open discussion to evaluate overall project progress, timeline, and the usefulness and potential impact of the project outputs. The project team will also check the success indicators of this project by tracking our project output dissemination. We will share the project information, some of the project outputs, and preliminary findings on our project website (see the Communication Plan) and will regularly conduct a web analysis to check the number of visitors, page views, and downloads of the outputs, including our conference presentations, proceedings, and/or journal articles. We will also communicate with attendees of our conference presentations to collect their feedback on our project.

Diversity Plan

We address diversity in our project in two specific ways. First, the composition of our project team (PIs and advisory board members) is diverse in terms of gender and cultural/ethnic backgrounds. We expect that having a diverse project team will be beneficial in bringing a variety of perspectives to bear on all of the phases of this project. We will also actively try to recruit data curators and data librarians to participate in our survey who work at historically black colleges and universities (HBCUs), colleges in the Appalachian region, women's colleges and universities, and/or Hispanic-serving institutions to investigate the capacity for big data curation at those institutions to ensure that their perspectives are taken into account. For our survey, we will sample all of the institutions listed on the HBCU Library Alliance website (http://www.hbculibraries.org), the Appalachian College Association website (http://www.acaweb.org), the Women's College Coalition website (http://www.womenscolleges.org), and the Hispanic Association of Colleges and Universities website

(http://www.hacu.net). These institutions are listed in the *American Library Directory* and each category will be weighted appropriately to ensure appropriate representation in our survey results.

Project Resources: Personnel, Time, and Budget

Project Team

Ayoung Yoon, Ph.D. is the **PI** and an Assistant Professor in the Department of Library and Information Science at Indiana University Indianapolis (IUPUI). She is also a RDA/US data share fellow. Her dissertation, *Data Reuse and Users' Trust Judgments: Toward Trusted Data Curation*, received the Eugene Garfield Doctoral Dissertation Award in 2015. Her research focuses on data curation, data reuse, and open data. She published a number of articles in top rated journals, such as Journal of the Association of Information Science and Technology (JASIST), College & Research Libraries (C&RL), and International Journal of Digital Curation (IJDC). Her previous work has been funded by Indiana University, University of North Carolina at Chapel Hill, and the Alfred P. Sloan Foundation.

Devan Ray Donaldson, Ph.D. is the **co-PI** and an Assistant Professor in the Department of Information and Library Science at Indiana University, Bloomington (IUB), where he directs specializations in digital curation and archives/records management. Donaldson is also Affiliated Faculty with the Data to Insight Center (D2I) at Indiana University. He is an internationally known digital curation researcher. His research interests include digital repositories, data sharing practices, mass digitization, preservation management, preservation metadata, trust, and security. His research appears in the *International Journal of Digital Curation* (IJDC), *Data Science Journal* (DSJ), and *Journal of the Association for Information Science and Technology* (JASIST). His research has been funded by the University of Michigan, Indiana University, the Alfred P. Sloan Foundation, and the United States Department of Energy. He holds a Ph.D. in Information from the University of Michigan.

Advisory Board Members

Our advisory board consists of 8 representative experts from the digital curation discipline from public and academic libraries (2 for each), curation technology (2), and the field of organizational capacity (2). Each member will bring his or her distinctive expertise and perspectives to our project.

Eben English, currently serves as a digital repository services developer at the Boston Public Library, and is responsible for Digital Commonwealth, Massachusetts's statewide digital library and service hub for the Digital Public Library of America. He has over ten years' experience building digital collections in academic and public libraries, is currently serving as co-principal investigator for the IMLS-funded "Historical Newspapers in Hydra: Building a Platform to Restore Access to Cultural Treasures" (LG-70-17-0043-17), and previously served as the principal investigator for the LSTA-funded "Voices of the Holocaust" project. His research interests include digital humanities, document markup languages, database design, the Semantic Web, and integrating open-source technologies into library services and collections.

Ann Hammond is the Executive Director at Lexington Public Library. She is a former scientist who brings her experience with data collection and analysis to her current position as executive director of the second-largest library system in the state of Kentucky. She uses data to identify the needs of her community and works with library staff to develop programs and services that have outcome-based goals. She continues to develop the organizational capacity of Lexington Public Library to respond nimbly and effectively to the identified opportunities within the greater Lexington community.

Robert McDonald is the Associate Dean for Research and Technology Strategies at the Indiana University Libraries. In his position he works to provide library information system services and discovery services to the entire IU system and manages projects related to scholarly communications, new model publishing, and technologies that enable the libraries to support teaching and learning for the IUB campus. In his role as Deputy Director of the Data to Insight Center, he works on new research related to large data analysis, storage and preservation through grant-funded and collaborative projects such as the HathiTrust Research Center. He also serves as the Data Steward for the IU Libraries. His research interests include digital libraries, discovery systems, technology management and integration of lean and agile frameworks, digital preservation, data preservation, learning eco-systems, storage systems, data cyberinfrastructure, and big data analytics.

Nancy McGovern is the Director for Digital Preservation at Massachusetts Institute of Technology Libraries. She is President of the Society of American Archivists (SAA), 2016-2017. She has thirty years of experience with preserving digital content, including senior positions at ICPSR; Cornell University Library; the Open Society Archives; and the Center for Electronic Records of the U.S. National Archives. She directs the Digital Preservation Management (DPM) workshop series; a well-regarded educational program offered more than fifty times since 2003.
Rodney Parker is an Associate Professor of Operations at the Kelley School of Business at IUB. He has previously been faculty at Yale University, Cornell University, and the University of Chicago. His research interests include how production and storage capacity limitations affect the inventory management of firms when acting optimally or under competition.
Tim Rogers serves as the Executive Director for the Metropolitan Library System of Oklahoma County, Oklahoma's largest public library. His interests in Big Data and its curation dovetail with his personal vision of the public library acting as the community's aggregator and archive of locally created government, business, education, and not-for-profit data. In addition to reducing collective barriers that stand in the way of improved educational, health, and commercial well-being, Tim believes that library managed access to public data will increase community trust, transparency, and accountability while also enabling library members to "learn smarter, work smarter, and live smarter."
Elaine Westbrooks is the Associate University Librarian for Research at the University of Michigan. In her position, she administrates the short and long-term objectives for the University of Michigan Library's support of the research enterprise. She provides operational leadership to the Library's subject liaison program as well as its copyright office and research data management (RDS) services. Her research interests include research data rights and metadata. She has co-edited *Metadata in Practice* with Diane Hillmann (2004) and *Academic Library Management: Case Studies* with Tammy Nickelson Dearie and Michael Meth (2017). She has presented her research at various conferences including the American Library Association, Coalition for Networked Information, Dublin Core, and the Association of College and Research Libraries.
Michael Witt is the head of the Distributed Data Curation Center and an Associate Professor of Library Science at Purdue University. His research explores the application of library science principles to the curation of research data and the development of new tools and practices to help librarians effectively steward data collections. As the co-chair of the Libraries for Research Data Interest Group of the Research Data Alliance and the co-lead of DataCite's re3data registry of research data repositories, he has worked to develop and promote international best practices for data librarianship.

Budget
The total amount requested from IMLS is $49,773 (Total direct costs: $31,602, Indirect costs (F&A @57.5%): $18,171). Although cost sharing is not required for the Planning Project, the School of Informatics and Computing at IUPUI and IUB will provide support in several ways, such as management of project finances, online survey tools, web spaces for project website, as well as physical spaces for the advisory board meeting. For full budget information see detailed **Budgets** and **Budget Justification**. The major project expenses are:

- Project personnel: $20,044 (PI and Co-PI summer salaries, Graduate Assistant ($13/h x 200h) with summer FICA)
- PI's Conference travel: $4,021 (Two domestic travels for conference presentations)
- Advisory board meeting: $7,242 (Travel to Indianapolis, hospitality)
- Other costs: $115 (Fabric research poster printing (36inx48in) for a conference presentation)

Time
This is one-year project. The **Schedule of Completion** lays out the duration of the major activities.

Communication Plan
The project team will utilize multiple channels to disseminate our findings and to communicate with audiences of the project as well as others who are interested. We will first create the project's website as a platform for

communicating, advertising our project, sharing our findings and outputs, and collecting any feedback from communities of interest. The project team will also present the project information, work-in-progress/preliminary findings, and final outputs at professional conferences, such as the International Digital Curation Conference (IDCC), RDAP, and PLA annual meeting. We will also attend academic conferences, such as the annual meeting of the Association of Science and Information Technology (ASIST), and submit journal articles for publication to reach out to our academic audiences. Those presentations and peer-reviewed publications will help us to evaluate and improve our work. All presentation and conference/journal papers will be linked to our project website for open access to the public, and the works will also be stored in the IUPUI ScholarWorks Repository to ensure long-term access and preservation.

Sustainability Plan

The findings from this planning project will be used to develop a framework for big data curation capacity in academic and public libraries. The project team will continue conducting research on this topic to disseminate our framework, to advance our framework, and to keep track of developments in technology related to big data. We expect that academic and public libraries will be able to integrate our framework into their conversations and planning for big data curation. Toward this end, we will market the framework broadly at conferences that staff members at academic and public libraries frequently attend (e.g., the ALA annual meeting). Additionally, based on the findings of our planning project, we will apply for the full Laura Bush 21st Century National Leadership Grant for Libraries to develop a toolkit that will enable academic and public libraries to assess their capacity for big data curation. The full project will help academic and public libraries become better prepared for big data curation as a result of a better understanding of their capacity for big data curation.

NATIONAL IMPACT

Our project will have a major impact on libraries, both academic and public. Despite agreement on the importance of having big data curation programs in libraries and recent efforts to build such programs, little previous effort has been made to systematically understand a library's capacity to perform curation work and build a sustainable program. A conceptual framework will assist library staff in understanding their current environments as well as potential impediments to building successful curation programs. Additionally, the framework will be used in a subsequent project to develop a standardized capacity assessment instrument, which will be tested, evaluated, and freely available to any academic or public library upon its full development.

In addition, our project can be seen as a complement to existing standards for Trustworthy Digital Repositories (TDRs) (e.g., *Trustworthy Digital Repositories: Audit and Certification* (TRAC) and ISO 16363). Such standards specify requirements for repositories regarding organizational infrastructure, digital object management, technical infrastructure, and security risk management. Academic and public libraries could use our framework and toolkit to assess their capacity for big data curation before attempting to become certified curators of those types of digital resources. In this manner, using our project outputs can help academic and public libraries become more prepared for TDR certification.

This project meets the IMLS agency-level goals by **supporting stewardship of an ever-increasing important part of library collections—data**. This includes facilitating the use and discovery of data to generate knowledge. Our project belongs to the category of **Curating Collections**; it will have a significant national impact on creating new services and/or improving existing services for the curation, preservation, and management of data across the country by assessing libraries' current capacity and by helping libraries identify areas where they need to expand their capabilities.

Schedule of Completion

Activities		2017	2018									
		Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
Literature review (Organizational capacity & Big data curation)	Framing questions & literature search												
	Literature assessment & analysis												
	Interpretation												
Online survey	Survey design, review & IRB												
	Data collection												
	Data analysis												
Framework development	Initial draft of framework												
Expert review	Focus group												
Final report & full project planning	Final framework												
	Final report writing												
	Full project planning												

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.

This project will produce datasets coming from surveys and focus groups. The datasets will be in the format of spreadsheets and text files. The PIs will hold the copyright of the datasets. The datasets can be reused for non-commercial purposes. The following Creative Commons license will be assigned: CC BY-NC. The dataset will be deposited in IUPUI ScholarWorks, which is IUPUI Institutional Repository (https://scholarworks.iupui.edu/).

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 copyright of the datasets will be held by the PIs. The PIs will allow the sharing of the datasets for non-commercial use. Creative Commons will be used to describe the conditions of access and use. The following license terms will be applied:

- Share - copy and redistribute the materials in any medium and format;
- Adapt - remix, transform, and build upon the material;
- Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use;
- NonCommercial - You may not use the material for commercial purposes (http://creativecommons.org/licenses/by-nc/4.0/)

The licensing information will be notified when any users attempt to download the dataset from the repository.

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.

N/A

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.

N/A

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.
N/A

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).

N/A

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).

N/A

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).

N/A

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).

N/A

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

N/A

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).

N/A

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).

N/A

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.

N/A

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.

N/A

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.

N/A

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.

N/A

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

N/A

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

N/A

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

N/A

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

N/A

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.

N/A

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

N/A

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

Name of publicly accessible source code repository: N/A

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 project will collect empirical research data through surveys and focus groups to achieve the research objectives. An online survey will be administered to public and academic libraries listed in the American Library Directory in order to understand current library data curation practices (Months 4-9). A focus group will be conducted to review and vet the framework with experts (Months 9-10), which will be developed from our systematic literature review (Months 1-3). The collected data will be analyzed quantitatively and qualitatively to accomplish the project goals.

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 data collected from surveys and focus groups will require approval from the Institutional Review Board (IRB) at Indiana University. We will submit IRB documents for surveys in April 2018 (Month 5) and for focus groups in August 2018 (Month 9). If both will be treated as exempt from IRB oversight, it will be approved within 2 weeks. Both PI and Co-PI have had multiple successfully approved IRB applications.

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).

The survey will collect demographic information. We will use identification numbers to ensure the anonymity of responses. Also, all responses will be aggregated in the analysis. Before releasing the collected data to the public, we will remove any information that might include personal information, and ID numbers will be used for each response.

In focus group transcripts, we will not disclose and de-identify any personally identifiable information (e.g., name of person and organization) from. Instead, we will use randomly generated numbers. All these processes will be approved by the Indiana University Office of Human Subjects prior to data collection.

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.

We will collect informed consent forms for surveys and focus groups. The consent forms will include the description of the project, participation process, confidentiality, compensation, and other information. In addition, the consent form will inform the participants the data sharing plan. Only those who agree to the consent form will be able to participant in the study.

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).

For collecting survey data, we will use the Surveymonkey under the IUPUI Department of Library and Information Science account, which provides an online survey template. The collected survey responses will be downloaded in the format of .CSV, and will be analyzed using the SPSS, a statistical analysis software program (https://www.ibm.com/us-en/marketplace/spss-statistics).

For the data from focus groups, we will create audio-visual files that will record the focus group dialogue in .mov format. Our student research assistant (hourly paid) will transcribe the audio into text files (e.g., .docx, .txt).

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?

A survey questionnaire and coding book for surveys; and a question document for focus group. Those documents will be stored and managed together with the associated data in the IUPUI ScholarWorks. Anyone who is interested in the dataset will be able to access those additional documents along with the datasets.

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

After the completion of the project, our survey and focus group data will be archived, managed, and disseminated through the in the IUPUI ScholarWorks. Relevant metadata, including keywords, will be added to make them easily searchable on the web. In addition, we will include the information about dataset access in our publications and will also share the links to the datasets on our project website.

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

Name of repository: IUPUI ScholarWorks

URL: https://scholarworks.iupui.edu/

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

Data management plan will be reviewed every time we create and store data. Specifically, we expect to review this plan when we complete collection of survey and focus group data. We will do a final review of our data management plan upon completion of this project.