RE-37-19-0082-19 - Trustees of Indiana University
Indiana University Bloomington

Abstract: Bridging the Gap between Scientists, Institutional Repositories and Data Management Practices

In this Early Career Development project, Dr. Devan Ray Donaldson from the Department of Information and Library Science at Indiana University Bloomington requests $330,408 from the Laura Bush 21st Century Librarian program for a three-year empirical investigation into the use of data repositories by scientists in order to address the following research questions: 1) What features do scientists think are necessary to include in data repository systems and services to help them implement the data sharing and preservation parts of their DMPs? 2) How do institutional, domain, domain-agnostic, and/or commercial data repositories compare in providing these features? 3) How do scientists use IRs during DMP implementation? 4) What barriers arise as scientists use IRs during DMP implementation? 5) How do IR staff interpret and respond to information about scientists' use of IRs? In answering these questions, the research project will contribute both theoretical and practical knowledge about scientists' data needs and practices in five domains where attitudes toward data sharing are currently evolving and shifting (e.g., atmospheric science, chemistry, computer science, ecology, and neuroscience). Outcomes of the investigation have the potential to inform best practices for librarians on decision support for: 1) what data repositories to recommend to researchers, 2) what features to add to IRs, and/or 3) whether IR infrastructure is necessary and cost effective.

In pursuit of the research questions, the Project Director (PD), Devan Ray Donaldson, with support from a graduate research assistant, will conduct three studies. First, the PD will conduct a focus groups study centering on the characteristics and features of data repositories that scientists from five domains think are important for FAIR sharing and data stewardship (Study #1). Building on these features and characteristics, the PD will conduct a second study to construct and assess a rubric for determining the appropriateness of data repositories that librarians can utilize during their data consultations with scientists from these fields (Study #2). Third, the PD will conduct a multiple-case study with a separate sample of comparable scientists to examine their creation and implementation of DMPs, including their use of IRs in this process (Study #3). From the case study findings, the PD will develop a decision support tool (e.g., a questionnaire with decision tree elements) to help librarians decide on what features to include in IRs, when and how to further invest in their IRs, or instead focus on directing scientists to other, alternative data repositories.

To address project performance goals, the PD will measure success based on two sets of activities. First, he will present and disseminate the rubric for determining the appropriateness of data repositories (from Study #1 & #2) to librarians who attend the Research Data Access and Preservation (RDAP) summit and Research Data Alliance (RDA) Libraries for Research Data Interest Group meetings. During these presentations, he will ask librarians to use the rubric in data consultations with scientists at their institutions. For those who express interest, he will collect their contact information and conduct follow-up interviews to assess how they used the rubric and gauge their perceptions of its usefulness. Second, the PD will present and disseminate the IR decision support tool (from Study #3) to librarians at Code4Lib and CNI meetings. During these presentations, he will ask librarians to apply the decision support tool to their IRs. For those who express interest, he will collect their contact information and conduct follow-up interviews to assess how they used the tool and gauge their perceptions of its usefulness.

The design, conduct, and interpretation of findings will be informed by an Advisory Board composed of digital curation experts. This research in service to practice centers on multiple key themes of the National Digital Platform, including preservation and infrastructure.

Indiana University Bloomington

Bridging the Gap between Scientists, Institutional Repositories and Data Management Practices

In this Early Career Development project, Dr. Devan Ray Donaldson from the Department of Information and Library Science at Indiana University Bloomington requests $330,408 from the Laura Bush 21st Century Librarian program for a three-year empirical investigation into the use of data repositories by scientists in order to address the following research questions: 1) What features do scientists think are necessary to include in data repository systems and services to help them implement the data sharing and preservation parts of their Data Management Plans (DMPs)? 2) How do institutional, domain, domain-agnostic, and/or commercial data repositories compare in providing these features? 3) How do scientists use institutional repositories (IRs) during DMP implementation? 4) What barriers arise as scientists use IRs during DMP implementation? 5) How do IR staff interpret and respond to information about scientists' use of IRs? In answering these questions, the research project will contribute both theoretical and practical knowledge about scientists' data needs and practices in five domains where attitudes toward data sharing are currently evolving and shifting (e.g., atmospheric science, chemistry, computer science, ecology, and neuroscience). Outcomes of the investigation have the potential to inform best practices for librarians on decision support for: 1) what data repositories to recommend to researchers, 2) what features to add to IRs, and/or 3) whether IR infrastructure is necessary and cost effective.

1. Statement of Broad Need

Open science by design (OSBD) is a vision where scholarly publications, the data that result from scholarly research, and the methodologies, including code or algorithms, that were used to generate those data are freely available and usable (National Academies of Sciences, Engineering, and Medicine, 2018). **There is currently a strong national need** to examine the relationship between OSBD as an ideal and what is necessary to make it a reality. One impediment to OSBD is the fact that scientists within and across disciplines differ in the nature of their research and practices surrounding treatment of data and code (National Academies of Sciences, Engineering, and Medicine, 2018). In order to address this issue, we need more research on the similarities and differences of scientists regarding management of their data, and we need more research on how to help scientists preserve and share their data despite these differences. Also, to help make OSBD a reality, scientists need to deposit their research data in one or more data repositories, with clear and persistent links among articles, data, and software to enable data sharing and preservation (National Academies of Sciences, Engineering, and Medicine, 2018). To address this issue, a better understanding of what scientists require of data repositories and services is necessary as well as a means of meaningfully comparing existing data repositories and services based on scientists' requirements so that they can select the appropriate places to deposit, publish, preserve, and share their data.

The Project Director (PD), Devan Ray Donaldson, has already conducted research on data management, including data sharing and digital repositories from the perspective of users (Donaldson et al., 2017; Fear & Donaldson, 2012) and repository staff (Donaldson & Bell, in press; Donaldson & Conway, 2010; Donaldson et al., 2016). **Building on his existing research agenda**, this project will focus on bridging the gap between OSBD as an ideal and OSBD as a reality. The PD will build and assess a rubric that librarians can use to help scientists compare data repositories and their services. The rubric will provide a mechanism for comparing various data repositories where scientists can publish their data based on the features that are important to them. The PD will also conduct a multiple-case study where scientists implement their DMPs, either using IRs or not. From the case study findings, the PD will develop a decision support tool to help librarians decide on what features to include in IRs, when and how to further invest in their IRs, or instead focus on directing scientists to other, alternative data repositories.

Digital curation is "the active management and enhancement of digital information assets for current and future use" (National Research Council, 2015, p. 10). Research data management (RDM) is an integral part of the research process and helps to ensure that scientists' data are properly organized, described, preserved, and shared. Recent mandates from federal funding agencies now require grant holders to make the data that they generate using grant funds publicly available. In particular, funding agencies are requiring scientists to submit

Data Management Plans (DMPs) as part of their grant applications. DMPs typically state what data will be created and how, and outline the plans for sharing and preservation, noting what is appropriate given the nature of the data and any restrictions that may need to be applied (Digital Curation Centre, n.d.). Scientists currently have several options for implementing DMPs, such as making data available on a publicly accessible web site, publishing data in a journal, or depositing data in a repository. Examples of repositories include: institutional repositories (e.g., DSpace@MIT, Massachusetts Institute of Technology's institutional repository), domain repositories (e.g., HEPData, a repository for publication-related high-energy physics data), data centers (e.g., NSIDC, National Snow and Ice Data Center), and domain-agnostic repositories (e.g., Dryad Data Repository). There are also commercial data repositories and services available (e.g., Figshare, Mendeley Data, Open Data on Amazon Web Services, etc.). Although these repositories and services differ in the platforms they use, their business models, their functionality, and their capacity for data curation, scientists could use any of them to comply with the requirements in their DMPs.

The growing number of choices about where scientists can publish their data have affected librarians' provision of research data services. For example, librarians are: 1) producing lists of data repositories, making them available on library websites, 2) consulting with scientists to help them find "the right" repositories for their data, and 3) expanding IRs or creating new IRs. Regarding the third point, research has shown that it is getting harder to justify the expense (e.g., in time, human, and computing resources) of providing infrastructure for research data in IRs, when few scientists are depositing their data in IRs, and many researchers are depositing their data elsewhere (Lowenberg, 2017). Research has also shown that, when data sharing and preservation are embedded in scientists' workflows and tied to journal article publication, scientists are far more likely to, for example, deposit their data anywhere their journal publishers or funders recommend or require (Lowenberg & Chodacki, 2018). These findings raise important questions about librarians' roles in supporting scientists with data management as well as how IRs should fit into this process. For example, should librarians preserve research data in IRs or create institutional data repositories? Should librarians retire or sunset their data repositories? Should librarians preserve and share some data in IRs and refer other scientists to share their data in alternative data repositories? Answers to these questions may involve gathering more information from scientists on their data needs and practices, including their requirements for data sharing, data discovery, and complying with funder mandates. Toward this end, librarians have conducted user assessments to guide the development of various research data services (e.g., Borghi, 2017; Carlson, 2010; Johnston, 2017; Wilson et al., 2010; Witt, 2008). However, more work still needs to be done.

1.1 Research on Scientists' Data Sharing Practices

Research on scientists' data sharing perceptions, attitudes, and practices seeks to measure the correspondence between what scientists say about data sharing and what they actually do. On average, scientists are more willing to share their data, and scientists actually do share their data more than they have in the past (Tenopir et al., 2011; Tenopir et al., 2015). Not only are scientists more willing to place at least some or all of their data in a data repository, some scientists actually do store their data in data repositories (Kim, 2017; Tenopir et al., 2015). Of course, this does not hold true for all scientists in all disciplines; some scientists are more disposed to sharing data than others. Thus, *there is a need to study varying individual cultures within the scientific community more closely, to understand how to continue to build infrastructure that promotes data sharing given the needs of different research communities* (Tenopir et al., 2015).

Scientists report the perceived amount of effort, time, and skill it takes to make their data shareable as barriers to data sharing (Kim, 2017; Tenopir et al., 2015). Scientists who deposit data in repositories were more likely to report that contributing data to a repository was either "difficult" or "very difficult" because of the perceived effort required to provide metadata during the submission process (Kowalcyzk, 2014). *These findings suggest a need for more research on these barriers and how to reduce them.*

1.2 Librarians, Data Repositories, and Research Data Services

Recent research on librarians who work on data repositories shows that they perform a wide variety of research data management services (Lee & Stvilia, 2017). Johnston (2017) outlines steps for curating research data in a data repository, several of which involve collaboration with and interaction between curators and scientists. For example, data curators need to conduct pre-deposit interviews with scientists to understand the

data, share concerns, and assess the long-term value of the data. As another example, data producers may deposit the data, but data curators may likely facilitate transfer of the data into the repository. Additionally, although scientists provide metadata for the data, data curators need to verify the metadata and review the available documentation. The data curator must also determine if the description is sufficient, and if not, seek out additional information from the scientist(s) or create additional documentation. This description of research data services and processes underscores the fact that there is an extensive process that must occur in order to make the plans that scientists specify in their DMPs regarding the use of data repositories a reality.

1.3 Research on Gathering Scientists' Requirements for RDM

Librarians who have conducted user requirements studies to guide the development of their research data services typically employ in-depth, qualitative assessments. For example, in a series of semi-structured interviews with scientists at the University of Oxford within the life sciences, medical physics, image analysis, and computing domains, Martinez-Uribe (2009) found that the scientists wanted: 1) advice on practical issues related to managing data across their life cycles, including help with data management plans, and assistance with data formatting; 2) secure storage for large datasets generated using high throughput instruments, and 3) sustainable and authenticated infrastructure that allows publication and long-term preservation of research data. In a series of semi-structured interviews with scientists at the University of Vermont within the physical and biological sciences, Berman (2017) found that some of the scientists were intrigued by the idea of having an institutional data repository, seeing a major benefit in not having to be responsible for the long-term management of their data if such a repository was in place. Other participants in the study shared their skepticism in whether an institutional data repository could meet their overall needs because they collect different types of data that have different requirements. Witt et al. (2009) conducted a series of semi-structured interviews with 19 researchers from 12 disciplines to better understand how the scientists handle, manage, and share their data and determine what information to collect about their data needs that are pertinent for curation. Witt et al. formalized their methodology into a Data Curation Profiles Toolkit to help librarians and other information professionals identify the data needs of scientists by interviewing them.

Adapting the Data Curation Profiles Toolkit into a focus group protocol, McLure et al. (2014) conducted five focus groups with 31 researchers who study national resources, natural sciences, medicine, engineering, and the liberal arts to discuss the nature of the datasets they create and maintain, how the researchers manage their datasets, and their needs for assistance and support in relation to sharing, curating, and preserving their datasets. Regarding their participants' views on library support, they were optimistic about libraries' role in enhancing the discoverability and dissemination of their data and research products. Some participants perceived the institutional data repository as facilitating not only data and document management but also accessibility and preservation. They expressed an interest in utilizing librarians' support in hosting their data and adding metadata to enrich their contextualization. The participants also made comments about their interest in subject-specific repositories and data dissemination facilitated by librarians. Participants' discussion during the focus groups provided critical insights into their interest in utilizing librarians and library resources regarding management of their data.

1.4 Scientists, Data Repositories, and DMPs

There is some empirical evidence that scientists intend to use data repositories as part of DMP implementation for their grant-funded projects. Parham and Doty's (2012) content analysis showed that roughly 14% of the DMPs that scientists at Georgia Tech submitted to NSF mentioned plans to utilize SMARTech, Georgia Tech's institutional repository, for long-term data sharing and archiving. Since their study, other studies have shown a slight increase in the number of scientists who mention that they will make use of data repositories in their DMPs as a means of sharing and preserving their data. For example, separate content analyses of scientists' NSF DMPs at the University of Illinois, the University of Michigan, the University of Minnesota, and Wayne State University have shown mention of these institutions' IRs over 20% (Bishoff & Johnston, 2015; Carlson, 2017; Mischo et al., 2014; Van Loon et al., 2017). These findings demonstrate that increasingly scientists are planning to utilize data repositories to help them implement their DMPs. However, *little research investigates whether the scientists actually follow through on these plans and what that process*

entails. This is important because recent research has found a number of disparities between scientists' intent to preserve their data and the actual stewardship of their data (York et al., 2018).

1.5 Research Gap

Even though librarians have taken great strides to better understand scientists and their data needs, and to use what they learn to inform their development of systems and services to support scientists with data sharing and preservation, **there is still more work that needs to be done**, because data sharing is still limited to a few fields, and practices within those fields are not well understood (Borgman, 2012; Borgman, 2015). Librarians need to be more integrated into scientists' research workflows to aid them in creating content that can be properly preserved and shared, while keeping in mind that scientists' needs "vary not only across disciplines but also within them" (Rudersdorf et al., 2018, p. 7). The answer is not to create a different system and service for each scientist to address their idiosyncrasies, but rather to continue to study scientists within and across domains to identify where the commonalities and differences are, and build on the commonalities. Now that there are new and different data repositories and services where scientists can share their data, *we need more research on how these data repositories and services can support scientists with implementing their DMPs*. Some fields have well-developed practices and infrastructure for sharing data. For example, in genomics fields, scientists have repositories for depositing specific genes, proteins and other structures (e.g., GenBank), and social scientists have well-established networks of repositories (e.g., ICPSR, Dataverse, etc.). One question is whether IRs can fill a niche need for the preservation and sharing of data from domains that generally lack an alternative such that librarians can mainly guide researchers to other existing services. Notwithstanding, **we currently lack an objective assessment tool which can compare the relative merits of the different data repositories and services that are available to help scientists with data sharing and preservation.** Since some librarians are currently considering creating institutional data repositories or expanding existing IRs to include research data, while other IRs are moving away from handling research data, *more research is needed on scientists' data practices **in the context of IRs** to inform decisions of when to use of IRs for handling research data, or when alternative data repositories are more appropriate.*

Existing research examines the content of DMPs for their mention of plans to use data repositories for data sharing and long-term preservation. However, research on this issue typically does little to assess scientists' follow through on the plans that they specify in DMPs. In practice, scientists may not utilize data repositories as they had planned for reasons we may (or may not) be aware of or understand.

The proposed project addresses these gaps by investigating scientists' requirements for data sharing and preservation as well as their requirements for data repositories and services. Additionally, this project addresses these gaps by observing and supporting scientists as they create and implement DMPs, preparing data for sharing, preservation, and deposit in IRs. **Building on the PD's existing research agenda**, *this project will contribute both theoretical and practical knowledge* about scientists' data needs and practices across five domains that are grappling with data sharing issues (e.g., atmospheric science, chemistry, computer science, ecology, and neuroscience). Outcomes of the investigation will provide librarians with a rubric and key insights on how to recommend data repositories to scientists from these domains for data management. Additionally, this project has the potential to provide librarians with a decision support tool and recommendations on whether they should create or expand IRs to handle research data, or abandon such efforts in favor of recommending scientists from these domains to other data repositories.

2. Project Design

To answer the research questions, this project includes three related studies. Table 1 maps the studies to the research questions, methods, and outcomes.

Research Questions	Study #	Methods	Outcomes
1) What features do scientists think are necessary to include in data repository systems and services to help them implement the data sharing and preservation parts of their DMPs?	1	Focus Groups	Data Repository Appropriateness Rubric Draft

2) How do institutional, domain, domain-agnostic, and/or commercial data repositories compare in providing these features?	2	Rubric development and testing	Rubric assessment
3) How do scientists use IRs during DMP implementation? 4) What barriers arise as scientists use IRs during DMP implementation? 5) How do IR staff interpret and respond to information about scientists' use of IRs?	3	Multiple-case study	Recommendations and decision support tool

Table 1. Mapping of project studies, research questions, methods, and outcomes.

2.1 Study #1 – Focus Groups Study (**Timeline:** Year 1)

During year 1, the PD will conduct five focus groups, one for each of five domains—atmospheric science, chemistry, computer science, ecology, and neuroscience—where the culture towards data sharing is shifting and evolving. **Note that IRB approval from the Indiana University Bloomington Office of Human Subjects will be obtained before any formal participant recruitment or data collection transpires**. The PD will travel to the major conferences of each field to recruit participants and conduct the focus groups on-site (e.g., American Geophysical Union Conference (atmospheric science); American Chemical Society Conference (chemistry); Symposium on Useable Privacy and Security (SOUPS'2019) (computer science); Society of Freshwater Science Conference (ecology); and Neuroscience 2019). For recruitment, the PD will attend sessions to learn about and meet potential participants; describe the study to conference attendees during breaks and social events; and publicize the study using the conferences' Twitter hashtags. The PD will distribute informed consent forms to those who express interest in participating based on these recruitment strategies. The PD will reserve a conference room at each conference to conduct the focus groups. For those who agree to participate, the PD will ask them questions about their data, past and present research projects, their data management, DMPs, what aspects of data management they would like help with, whether they think libraries can help, and data sharing; the questions are based on a previously developed and empirically-tested focus group protocol (McLure et al., 2014). The PD will use these questions to lead into a discussion about the FAIR principles and what features they think are necessary to include in data repository systems and services to help them implement the data sharing and preservation parts of their DMPs. *The PD will also ask about participants' expectations for: file size acceptance, licensing, embargo periods, data discoverability, and reuse.* The focus groups will be video-recorded and transcribed. After conducting the focus groups, the PD, Graduate Research Assistant (GRA), and an IU-Undergraduate Research Opportunities in Computing (UROC) student **will perform thematic analysis of the focus group data following these steps**: 1) familiarizing ourselves with the data by reading the transcripts and watching the video recordings at least twice, 2) generating initial codes based on patterns of consensus that we begin to recognize, 3) searching for themes, 4) reviewing themes, 5) defining and naming themes, and 6) producing a report of our findings (Braun & Clarke, 2006). Themes pertaining to scientists' expectations and desired features of data repositories for FAIR data sharing and preservation will be coded using the Coding Analysis Toolkit (CAT).[1] The PD, GSA, and UROC student will code each focus group transcript, calculating inter-reliability in CAT until we reach a Cohen's kappa coefficient of 0.40, the lower-bound of acceptability for kappa according to Banerjee et al. (1999).

2.2 Study #2 – Rubric Study (**Timeline:** Year 1-2)

The PD will use the themes from the focus groups study to produce an analytic rubric for data repository appropriateness for scientists. The PD will follow procedures in Brookhart (2013) for producing and testing analytic rubrics, including: developing the rubric tool, testing the rubric tool, training evaluators on rubric use, collecting data, and analyzing and reporting the data. The features and requirements that the scientists mention during Study #1 will form the dimensions of the analytic rubric. To test the rubric, 15 scientists comparable to those who participated in Study #1 will examine three institutional repositories, three domain data repositories, three domain-agnostic data repositories, and three commercial data repositories to assess the extent to which

[1] CAT is a free, open-source qualitative data analysis tool. http://cat.texifter.com

each has the features and requirements listed in the rubric. This testing will generate empirical data (i.e., ratings) on whether these different types of data repositories possess the features and requirements that scientists desire. Based on examples and guidance in Brookhart (2013), Table 2 illustrates the possible structure of the analytic rubric.

	1 – Does not meet requirements	**2 – Meets requirements**	**3 – Exceeds requirements**	**Comments**
Feature #1 Score:				
Feature #2 Score:				
Feature #3 Score:				
Total:				

Table 2. Data Repository Appropriateness Rubric Structure Example.

Since the intended audience for this rubric is librarians, to help them advise scientists on which data repositories and services to deposit their data, **the PD will disseminate the rubric to practitioners in presentations** at the Research Data Access and Preservation Association (RDAP) Summit meeting and the Libraries for Research Data Interest Group (IG) meeting at the Research Data Alliance (RDA) plenary in Spring 2020. Hundreds of librarians who work on research data services regularly attend these meetings. **To evaluate the rubric with his intended audience**, the PD will ask the librarians (at least 5) who attend his presentations to begin using the rubric during data management consultations with scientists at their home institutions and report on their experience at subsequent RDAP and RDA meetings in Fall 2020 and Spring 2021 respectively. **IRB approval from the Indiana University Bloomington Office of Human Subjects will be obtained before any formal participant recruitment or data collection transpires**.

Because this study is at the proposal stage, it is impossible to know exactly what features will be included in the rubric, or if there will be five rubrics, one for each discipline, instead of one rubric. This will all depend on analysis of the focus groups data, and how many similarities and differences are identified within and across the five scientific disciplines under investigation.

2.3 Study #3 – Multiple-Case Study (**Timeline:** Year 1-3)

To address the last three research questions, this project includes a multiple-case study, with the following five components: 1) a case study's questions, 2) its propositions (if any), 3) its unit(s) of analysis, 4) the logic linking the data to the propositions, and 5) the criteria for interpreting the findings. This case study, using Yin's (2017) typology, is classified as an embedded, multiple-case design with descriptive and exploratory elements. The unit of analysis in the case study is the scientists (n=25). This case study is embedded because it has multiple elements (e.g., scientists' artifacts: DMPs, grant proposals; interviews, observation, institutions, IRs, deposit policies, funding agency requirements, IR staff, etc.). It is multiple-case because the PD will investigate multiple scientists across five different domains and institutions. *Note, these scientists will be comparative in research interests and disciplines to those who participate in Study #1 and #2.* The case study is descriptive in the sense that it will document the steps that transpire as scientists create and implement DMPs. It is exploratory in the sense that it will explore how the scientists make use of IRs during DMP implementation as well as explore IR staff's perceptions of and responses to any of the scientists' barriers to use of IRs during DMP implementation.

The context of the unit of analysis is the institutions where the scientists are employed and their corresponding IRs, both of which may affect how the scientists can (or cannot) make use of IRs during DMP implementation. The criteria for selecting the cases are scientists who are actively engaged in research. The IRs at Institution A, B, and C accept data; the IRs at Institution D and E currently do not. *This research design provides an important opportunity to meaningfully compare the data practices and DMP implementation strategies of scientists across five domains where data deposit in IRs is and is not available to assess what impact this does (or does not) have on the scientists' data management.* Table 3 illustrates the structure of the

multiple-case study design. Altogether, this case study will include six elements: 1) the case study report; 2) the case study protocol, 3) the code book, 4) the evidence tables and reliability reports, 5) the description of items in the case study database, and 6) the case study database.

Domains	**Institution A IR-A**	**Institution B IR-B**	**Institution C IR-C**	**Institution D IR-D**	**Institution E IR-E**
Atmospheric Science (AT)	Scientist A-AT	Scientist B-AT	Scientist C-AT	Scientist D-AT	Scientist E-AT
Chemistry (CH)	Scientist A-CH	Scientist B-CH	Scientist C-CH	Scientist D-CH	Scientist E-CH
Computer Science (CS)	Scientist A-CS	Scientist B-CS	Scientist C-CS	Scientist D-CS	Scientist E-CS
Ecology (EC)	Scientist A-EC	Scientist B-EC	Scientist C-EC	Scientist D-EC	Scientist E-EC
Neuroscience (NS)	Scientist A-NS	Scientist B-NS	Scientist C-NS	Scientist D-NS	Scientist E-NS

Table 3. Multiple-Case Design for Study #3.

The PD will collect multiple sources of evidence in this case study. The following lists the data the PD plans to collect and the study propositions to which they correspond:

Study Proposition #1 and #2

Libraries that provide research data services, including IRs that allow data deposit, data consultation, and DMP consultation, will be able to accommodate scientists who have never used IRs for data deposit, long-term preservation, and sharing as they make use of them during DMP implementation (Study Proposition #1). Notwithstanding, challenges may arise during this process, both technical and social (Study Proposition #2).

Data for Study Proposition #1 and #2

- Recorded video and audio; observational data: Direct observation of scientists, i.e., each case, as they use IRs (e.g., as they deposit their data into IRs) (either in person or via recorded zoom videoconference sessions where the scientists will share their screens with the LIS graduate students and think aloud as they deposit their data into their respective IRs). This process will occur for data generated for at least one study related to at least one DMP for each scientist (i.e., at least 15 times). The audio from all observations will be transcribed.
- Scientists' DMPs.
- Scientists' grant proposals.
- Field notes: LIS students will take field notes when they meet with the scientists as they deposit data into IRs, either in-person or virtually.
- Research data services information from libraries' websites.
- Data deposit guidelines from IRs' websites.
- Data preservation and sharing information from IRs' websites
- Interviews: Of scientists, i.e., each "case," to understand their experience using IRs and to find out what barriers, if any, they thought impacted their ability to use IRs to implement their DMPs (e.g., interviews, conducted either in person or via recorded zoom videoconference sessions, after data deposit to discuss their experience, any barriers and how they overcame them). The audio from all interviews will be transcribed.

Study Proposition #3 and #4

Scientists may decide not to use the IRs or solicit the help of librarians or LIS students (Study Proposition #3). In turn, IR staff or LIS students may support scientists' data management needs by helping them make use of IRs or referring them to other repositories (Study Proposition #4).

Data for Study Proposition #3 and #4

- Recorded video and audio; observational data: Direct observation of scientists as they consult with IR staff for support in using the IRs (e.g., for data deposit, DMP creation or evaluation) (either in person or via recorded zoom videoconference sessions). The audio from all consultations will be transcribed.
- Email: Any forwarded messages between scientists and IR staff as they ask and answer questions related to use of IRs.

- Recorded video and audio; observational data: For example, direct observation of LIS students as they help scientists create and implement DMPs (e.g., help with deposit data, help with creating the requisite metadata for the scientists' datasets pre-deposit, etc.) (either in person or via recorded zoom videoconference sessions). The audio from all sessions will be transcribed.
- Email: Any forwarded messages between scientists and LIS students as they ask and answer questions related to use of IRs for DMP implementation.
- Interviews: Of IR staff to understand their interpretation of and response to information about the scientists' use of IRs (conducted either in person or via recorded zoom videoconference sessions). The audio for all three interviews, one for each IR manager, will be transcribed.

LIS students who take the PD's digital curation courses will actively participate in data collection for this project to learn about scientists' data practices across domains and practice providing research data services to scientists. Approximately 15 students enroll in the PD's course each year. Consequently, every student will be assigned to at least one scientist, i.e., one case, and the majority of the class will be assigned to two scientists, i.e., two cases. They will either analyze one of their scientists' existing DMPs, or help their scientists create DMPs for new projects, depending on the scientists' preferences. Examples of LIS students' data collection activities include: 1) observing and/or supporting scientists as they implement their DMPs, 2) observing and/or supporting scientists as they use IRs, 3) recording these actions using zoom videoconferencing and screen sharing software, 4) taking field notes of these activities, and 4) supporting scientists with DMP creation.

All data will be compiled into a case study database in a password-protected, dual-authenticated environment on a secure server and backup server at IU. After data collection, all video and audio recordings will be destroyed. All transcripts will keep the institutions, IRs, and study participants anonymous to protect human subjects' confidentiality. **IRB approval from the Indiana University Bloomington Office of Human Subjects will be obtained before any formal participant recruitment or data collection transpires**.

For the observation transcripts and interviews of the scientists and IR staff, the unit of analysis is the scientists' and IR staff's statements. For email and IR website information, the unit of analysis is statements about IR use. For the DMPs and scientists' grant proposals, the unit of analysis is any statement about data sharing or preservation or any statement about the use of an IR for data preservation or sharing.

In the absence of an a priori set of codes, an emergent coding technique will be used to generate the initial codes (Neuendorf, 2017). The codes will be refined until the PD, GRA, and UROC student achieve a Cohen's kappa value greater than 0.40 for each code. This process will help to generate the code book for the project.

Next, the PD will go through the codes again and code each unit of analysis. Final reliability will be established by taking a random sample of 50 of the units of analysis and having the GRA determine which code(s) belonged to each of the 50 units of analysis. The GRA's ratings will be compared to the ratings of the PD on those 50 units of analysis using percent of overall agreement and kappa. Final interrater reliability results will be calculated. After final coding of the data, the PD and GRA will create evidence tables that indicate which portions of the documents in the case study database supported particular codes. After coding all of the data and collecting evidence for each finding, we will piece together the findings to assemble a decision support tool (i.e., a questionnaire with decision tree elements) with two potential applications for librarians who plan to support researchers from the five domains under investigation: 1) support for deciding what features to add to IRs, and 2) support for deciding whether IR infrastructure is necessary and cost-effective.

This case study addresses construct validity by: using multiple sources of evidence; establishing a chain of evidence (i.e., creating evidence tables); and having key informants (i.e., IR staff) review drafts of the case study report. The Advisory Board will also give feedback throughout the project. This case study addresses internal validity with pattern-matching, and it addresses external validity by using replication logic in a multiple-case study. And finally, this study addresses reliability by using a case study protocol, developing a case study database, and examining interrater agreement for each of the codes in the code book.

Since the intended audience for the case study is librarians, The PD will disseminate the case study findings and decision support tool to practitioners at Lib4Code and CNI breakout sessions during Year 3 of the project to get feedback and encourage adoption.

2.4 Dissemination Plan

In addition to disseminating the project findings and outcomes from each study in the ways described above, outputs of the project will include the publication of results in scholarly journal articles and peer-reviewed conference proceedings (e.g., *International Digital Curation Conference, Data Science Journal, International Journal of Digital Curation, ASIST Annual Meeting, etc.*). The PD will also disseminate project findings annually at the Indiana Library Federation (ILF) conference, in ACRL virtual conference webcasts as well as a final report to IMLS.

2.5 Performance Goals and Outcomes

The PD will measure success of the project based on two sets of activities. First, he will present and disseminate the rubric for determining the appropriateness of data repositories (from Study #1 and #2) to librarians who attend the Research Data Access and Preservation (RDAP) summit and Research Data Alliance (RDA) Libraries for Research Data Interest Group (IG) meetings. During these presentations, he will ask librarians to use the rubric in data consultations with scientists at their institutions. For those who express interest, he will collect their contact information and conduct follow-up interviews to assess how they used the rubric and gauge their perceptions of its usefulness. Second, the PD will present and disseminate the IR decision support tool (from Study #3) to librarians at Code4Lib and CNI meetings. During these presentations, he will ask librarians to apply the decision support tool to their IRs. For those who express interest, he will collect their contact information and conduct follow-up interviews to assess how they used the tool and gauge their perceptions of its usefulness. These measures of success will allow the PD to gauge the usefulness of project outputs by his intended audience.

2.6 On confidentiality of study sites and participants

IMLS full proposals and supporting documents can be subject to FOIA requests, as such the PD has decided not to name or provide details of the institutions, IR staff, or scientists who have committed to this project. For brief descriptions of proposed field sites please reference the original Preliminary Proposal.

2.7 Project Risks and Risk Mitigation Strategies

The major threat to each of the studies in this project is subject participation. For study #1, scientists may not be interested in participating in focus groups because they will be busy with conference activities. For study #2, librarians may not be interested in using the rubric or they may not be able to find scientists to practice using it with to evaluate its intended purpose. For study #3, although informally the PD has reached out to scientists who are interested in the project, when it is time to conduct the project and follow formal participant recruitment strategies, they may no longer have an interest in participating in this study. And even if they are interested, over the three-year project period, the scientists could retire, be promoted to more administrative positions that pull them away from their research, not get tenure, or, for a host of personal reasons, may not be able to sustain a level of participation in this project that is necessary for its success. To mitigate these risks, I plan to provide incentives ($50 amazon gift cards) for the focus group participants in study #1, the librarians and scientists in study #2, and the scientists in study #3. Also, for study #3, to compensate the IR staff for their time in supporting the scientists and for their time in participating in interviews and reviewing the project findings, the IR staff will be compensated for their time commensurate with their hourly pay.

2.8 Research Team

Project Director

Devan Ray Donaldson, PhD, is an Assistant Professor in the Department of Information and Library Science at Indiana University Bloomington where he directs a specialization in Digital Curation. He has published more than fifteen articles on a broad range of issues regarding research data management, digital repositories, and their users. He is a recipient of the 2017-2018 Indiana University Trustees Teaching Award.

Graduate Research Assistant

To be recruited for assistance in Years 1-3 of the project. The roles and responsibilities of the GRA will include active participation in data collection, processing the data (e.g., sorting; labeling; organizing files;

anonymizing identities; and depositing data to secure storage); active participation in data analysis, including coding and working with the PD on collaborative writing, and presenting the project's findings at conferences. For job description, see Resumes of Key Project Staff document.

Advisory Board (AB)

Six experts have agreed to serve as pro bono advisors: Jake Carlson (University of Michigan), John Chodacki (California Digital Library), Margaret Hedstrom (University of Michigan), Simon Hodson (CODATA), Clifford Lynch (Coalition for Networked Information), and Carol Tenopir (University of Tennessee); they all have expertise in digital curation, IRs, and research data management. **The advisers will provide advice, ongoing assessment, and evaluation as necessary.** The AB and PD will meet for three video-conference call meetings annually throughout the project **as a method of evaluating data collection and progress towards sharing findings.** The AB *will evaluate the design and progress of the project, assuring the quality and dissemination of conclusions.* As needed, the PD will consult with each advisor individually based on their knowledge and expertise. The AB has also agreed to provide course correction advice if the project falls off track or if threats to the project arise that require additional risk mitigation strategies beyond what the PD has already planned for.

3. Diversity Plan

This research project addresses diversity issues in four specific ways. First, the PD will study the data management practices of scientists from a broad range of institutions and disciplines (e.g., atmospheric sciences, chemistry, computer science, ecology, and neuroscience). Since prior research has shown that data sharing and management practices can vary by discipline, incorporating this type of diversity into the project is vital and will meaningfully contribute to the project outcomes. Second, for all three studies, the PD will actively seek out scientists and librarians from different racial/ethnic backgrounds and genders to participate in the project to incorporate racial and gender diversity. Third, for study #3, the project incorporates a diversity of three different IRs, each using different software platforms to offer a diversity of technical infrastructure to the project. And finally, The PD will collaborate with underrepresented minority students in the IU-Undergraduate Research Opportunities in Computing (UROC) program and summer research students from Historically Black Colleges and Universities (HBCUs) and other Minority Serving Institutions (MSIs) through the IU-MSI STEM Initiative. In the past, the PD has worked closely with UROC students to provide them with research experience on how to conduct literature reviews as well as how to collect and analyze data and prepare study findings for presentation. Throughout the project, the PD will involve UROC students so they can get first-hand experience on how to collect and analyze data related to data management research.

4. Broad Impact

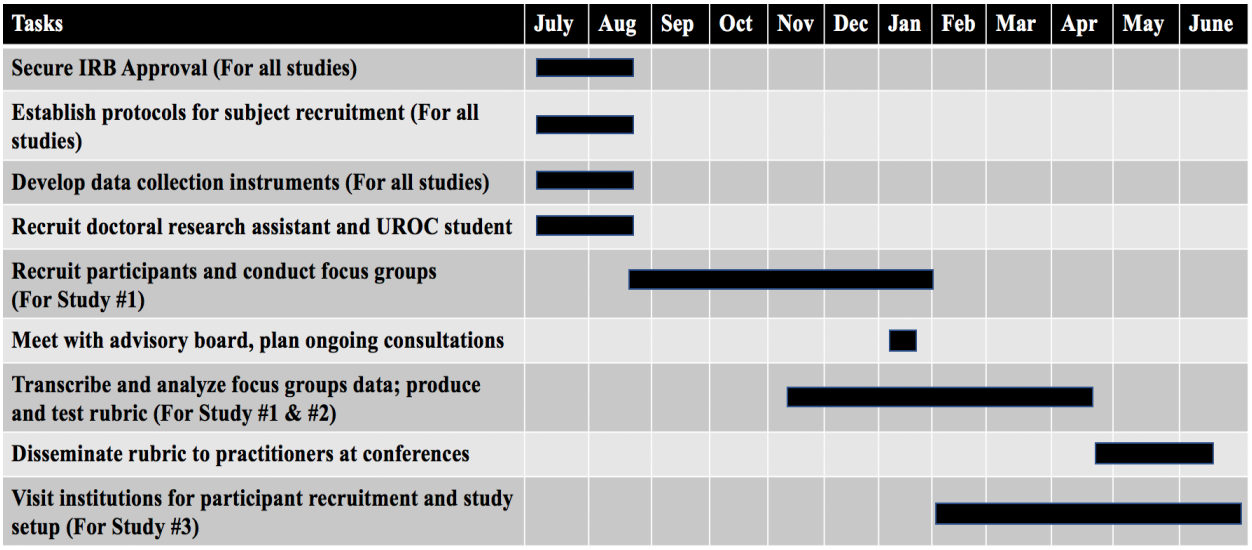
This research in service to practice centers on multiple key themes of the National Digital Platform, including preservation and infrastructure. Regarding the preservation theme, this project will develop, test, and disseminate to a large group of librarians who currently work on research data services a Data Repository Appropriateness Rubric to aid them in helping scientists select appropriate places to deposit and share their data. Regarding the infrastructure theme, this project will produce and disseminate a set of recommendations and a decision support tool to librarians about their capacity to address scientists' data sharing and preservation needs with existing infrastructure (i.e., IRs). Overall, this project will provide crucial data for steering resources for research data infrastructure to areas of critical need.

Additionally, **this project has a number of educational benefits**. LIS students who take the PD's digital curation and data management courses, most of whom have backgrounds in the humanities or the social sciences, will have the opportunity to learn from scientists about their domains, their data, and how they plan to manage them; the students will also practice providing research data services support to these scientists as they create and implement DMPs. Such opportunities will provide the students with hands-on educational experiences and will balance theory and practice, all of which are essential for positioning LIS graduate programs for 21st century practice (Sands et al., 2018).

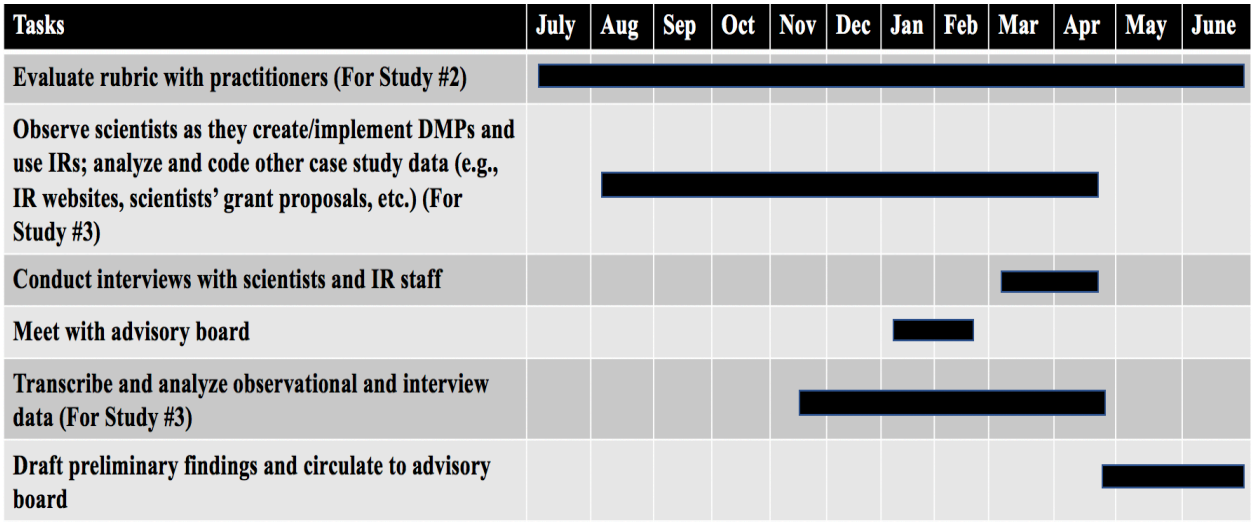
Indiana University Bloomington

Schedule of Completion

Year 1: July 2019 – June 2020



Year 2: July 2020 – June 2021



Indiana University Bloomington

Year 3: July 2021 – June 2022

Tasks	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
Observe scientists as they implement DMPs and use IRs (For Study #3)												
Analyze and process data collected, conduct inductive and deductive analysis; produce decision support tool												
Disseminate decision support tool to practitioners for feedback and evaluation												
Meet with advisory board												
Draft and submit conference papers, journal articles, and reports for publications												
Present findings at professional conferences												
Prepare and submit final report to IMLS												
Anonymize data and deposit in appropriate repositories												

DIGITAL PRODUCT FORM

Introduction

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

Instructions

All applications must include a Digital Product Form.

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

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

Part I: Intellectual Property Rights and Permissions

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

The main research products are a rubric for evaluating data repositories and a decision support tool for insitutional repositories. Both will be published under a Creative Commons Attribution 4.0 license, which will allow librarians to access and build on them. Other research products will appear as text and images in peer-reviewed journal articles, conference proceedings, and a final report. Primary source data will include transcripts from interviews, and observational notes in order to facilitate reuse. The PD will own the intellectual property of the dataset and resulting publications. Preprint copies of all published research and the final dataset will be submitted to the IU ScholarWorks repository (which is open access) in accordance to current best practice guidelines for preservation, future access, and reuse.

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.

Indiana University Bloomington will impose no condition on access and use of the published research made available through IU ScholarWorks. Primary data and other supporting materials created or gathered in the course of the research project will be shared with other researchers and LIS practitioners upon reasonable request. Access to research data will follow a set of access protocols guided by the standards of Indiana University's Institutional Review Board in order to the protect the privacy and confidentiality of study participants. All products will include attribution of the investigators and the funder, IMLS for supporting the research collecting and generating of data.

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.

Investigators from the project will publish the results of their work, which may include a selection of the data. Papers will primarily be published in peer-reviewed journals and conference proceedings. These data will be generated from focus groups, observation, rubric ratings, and semi-structured interviews with scientists from five domains and IR staff. Qualitative data (focus groups, interviews, transcripts, field notes) may involve privacy concerns. Personal information about participants will by anonymized and in some cases not collected to maintain research subjects' privacy. As such, transcripts will be de-identified and transcripts will not be published to reduce the possibility of disclosure or identification of persons or institutions. The PD will publish coding schema resulting from inductive analysis making anonymized data available to researchers if requested.

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

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

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

Digital content include: a rubric, decision support tool (i.e., a questionnaire with decision tree elments), published journal articles, and conference proceedings. Each of these resources will be published as PDFs and downloadable documents. Transcripts and codebooks will be formatted as PDFs or .docx word documents. Video recordings of focus groups (n=5) and zoom videoconference screen sharing observation (n=25) will be .mp4.

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.

The PD and research assistant will use machines and software provided by Indiana University, in addition to encrypted cloud storage provided by the university. Transcription services will be contracted later at a competitive market rate.

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).

File formats will include .docx, .pdf, .mp4, and .pptx

B. Workflow and Asset Maintenance/Preservation

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

The PD will manage the proposed deliverables, review all data and data project products before publication. The graduate research assistant will also review products as they are synthesized, and participate in quality assurance of the coding, transcription, and anonymization of collected materials.

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).

IU ScholarWorks will address preservation and maintainence of the digital assets after the award period of performance.

C. Metadata

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

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

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

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).

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.

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 project will collect and transcribe focus groups data on scientists' perceptions of features and characteristics for data repositories (from Study #1: Yr. 1), scientists' ratings of those features and characteristics (from Study #2: Yr. 1), observational data of scientists as they implement their DMPs and use IRs (from Study #3: Yr. 1-3), interview data from scientists and IR staff about scientists' experience using IRs (from Study #3: Yr. 2-3), and interview data from librarians who test the rubric and decision support tool that will be created in this study as a means of measuring the performance goals (Yr. 3). The overall reason for collecting these data is to provide a foundation for tools that librarians can use to provide research data services to scientists and make better informed decisions regarding RDM.

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?

Yes, IRB approval is required and all data collection practices will be reviewed by the Indiana University's IRB. An IRB application will be submitted after grant approval, as outlined in the schedule of completion, I expect to secure approval in the beginning of Year 1, with reapplications for approval as directed by the 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).

The video recordings will record personally identifiable information from interview subjects. After the focus groups and observational sessions, the recordings will be transcibed and the video will be destroyed. Every effort will be made in compliance with IU's IRB research protocols to preserve confidential information of the scientists or institutions. Final datasets submitted for public release will be anonymized and contain no personally identifiable information.

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.

Researchers will obtain verbal consent according to a human subjects' protocol and such verbal consent for participation will be recorded before the interviews. Once interviews have been transcribed, the audio recordings will be deleted to preserve identifiable voice information. During the implementation of the research project, associated research data will be backed up on a password-protected secure server maintained by Indiana University's computing infrastructure, in order to protect from loss of data from hardware failures, fire, theft, etc.

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).

Investigators will use notebooks, computers, zoom video conference software, word processing software, and qualitative transcription software to generate and capture data from the project. These data will be converted to open formats (e.g., ODF, PDF, JPEG) for preservation and future re-use. Interviews will be transcribed and subsequent coding schema, reports, articles, and conference proceedings will be generated as .pdf, .docx, and .csv. In Year 3 anonymized interview data and accompanying codebook will be included in the transcription data deposited in the IU ScholarWorks institutional repository.

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?

The codebook and interview transcripts will be stored as .docx documents.

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

Upon completion of the project, final dataset (including documentation, codebooks, and proper citation) will be submitted to IU ScholarWorks, a trusted, public and authenticated data repository.

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

Name of repository:

IU ScholarWorks

URL:

https://scholarworks.iu.edu

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

A long-term strategy for the maintenance, curation, and archiving of data will be implemented. The advisory board will monitor the implementation, reviewing the data management plan three times during annual meetings. The PD will review the data management plan every 6 months throughout the course of the project. Datasets will be curated and preserved according to the practices of IU ScholarWorks.