Extending data curation to interdisciplinary and highly collaborative research

Statement of National Need

This proposal addresses the challenge of curating data that is produced during interdisciplinary and highly collaborative research (IHCR), defined as research that integrates resources and expertise across disciplines and institutional settings. IHCR is a recognized priority in addressing complex research problems and providing solutions to societal challenges [1], [2]. Such research, however, places higher demands on communication, learning, and trust, all of which can have implications for how associated data is collected, managed, and preserved [3]. Academic libraries began to develop services in support of such research [4], but the increasing number of interdisciplinary collaborations and the unique challenges of working with IHCR data highlight a **national imperative to develop curation models aligned with researchers' needs that fully leverage both libraries' expertise and library-technology-research partnerships**.

The landscape of research data services in libraries is diverse. Many libraries offer basic data services such as data management plan (DMP) assistance and data consultations, while others have moved into more advanced services, such as supporting collaborations, data analysis and sharing, and submission of data to centralized interdisciplinary repositories [5], [6]. Effective attempts to develop services in support of curation recognize the need to engage with researchers as early as possible in the research lifecycle, turning proposal development and associated data management plans (DMPs) into important mechanisms to connect with researchers and address their data needs [7].

While a significant amount of effort has been devoted to the development of technical capabilities for interdisciplinary data curation [8]–[10], studies of interdisciplinary science show an unmet need for both technical and cultural adaptation of research practices to the contexts of interdisciplinarity. Paths to interdisciplinary and collaborative research build upon interdependency and information sharing as researchers mobilize and accumulate necessary resources and learn to accommodate different research cultures [3]. Collaborative practices include exchanges of facts, processes, methods, ideas, and technology [11]. Through co-presence, comparison, and sharing, interdisciplinary programs move from independent co-existence to connected networks [12]. Sound documentation, flexible tools, and strong relationships among data managers, data creators, and data users are key factors in supporting robust collaborations [13], [14].

As a developing practice, data curation in academic institutions has mostly focused on supporting disciplinary communities, with IHCR data curation existing mainly within specialized projects or centers. IHCR teams will benefit from increased awareness of the available tools and from tailored support that connects their data and research needs with existing infrastructure and data management professionals. To summarize, the significant national need for IHCR data curation stems from the needs of several stakeholders: 1) data producers who face obstacles in overcoming their differences and creating productive environments; 2) librarians and data managers who seek to provide useful services fitting in with what researchers are already doing; 3) data consumers who need data that they can trust and use in their own context; and 4) the public, which needs research data with the potential to solve difficult problems to be both broadly available and empowering.

This proposal addresses this multi-stakeholder need by studying IHCR data practices, expanding existing library services, and connecting researchers with curation and cyberinfrastructure expertise.

We consider this project to be in the *pilot stage* of maturity moving into the *scaling* phase, understanding scaling as both scaling into multiple institutions and scaling (expanding) the existing services within each institution. We will test our approaches with nine use cases, offer them for testing and peer review within the broader community, and identify ways in which to incorporate adaptations and feedback. Previous research has identified this area as an area of challenge and discussed user needs. We have identified use cases with which we will engage deeply to generate evidence of how the proposed solutions and best practices can be implemented and adopted. Additionally, our training and peer evaluation events will build the foundation for scaling and further adoption.

The proposed project will offer a theoretical understanding of IHCR data practices and provide insights into the tools and processes that interdisciplinary researchers employ in their work. It will use the knowledge derived from studying IHCR practices to develop practical models of embedded curation and to make recommendations about how library and research technology units can engage with IHCR teams, therefore contributing to the *Research in Service to Practice* category.

Project Design

The proposed project will examine IHCR practices and address the following research questions:

RQ1) What are the patterns in data curation in IHCR?
RQ2) How do IHCR teams make decisions and negotiate differences in their data work throughout its lifecycle?
RQ3) What is the role of data management professionals in supporting IHCR data curation and cultures of interdisciplinarity?

To address these questions, we will use the case study approach that includes theory-informed sampling, multiple data collection methods, intra- and cross-case data analysis, and iterative examination of evidence for theoretical insights [15]. The case study approach will allow us to examine a complex phenomenon in its own environment while taking into account the conditions of institutional settings that affect the behavior of the researchers involved in the study.

To build our pool of use cases, we invited interdisciplinary and collaborative teams across three campuses to participate in our study. The selection of use cases was informed by the concept of “purposeful curation,” which emphasizes scalability and the needs of information users in provisioning library services [16], [17]. With its emphasis on the commitment to users and building systems that work in concert with existing knowledge infrastructures while envisioning future uses of data, purposeful curation aims to address the community’s core concerns, to prioritize local services, and to account for scale and granularity. With these guidelines in mind, we selected projects that are actively collecting or analyzing data, cover two or more disciplines, types of data, methodologies and institutions or units, and have the potential to make an impact outside of their own institution (e.g., in the form of data sharing, policy-making, education, etc.). The following nine use cases were selected[1]:

1. **Sustainable ecology, agriculture and food:** Bringing together scholars from cognitive science, anthropology, public health, and history, the team of about 25 researchers, students, and educators works with communities around the world to study the systems of food supply, processing, distribution, and consumption.

[1] Names and exact details of each case are omitted from the proposal to preserve the anonymity of study participants.

2. **African slavery digital collection:** Historians, computer scientists, and geographers have created a public digital archive of images of African slaves and their descendents in the Americas and are currently working on developing a geospatial component.
3. **Specialty crops research initiative:** This large project is a collaboration between researchers in genetics, entomology, and economics to use diverse data streams in development of domestic varieties of the vegetable soybean edamame.
4. **Disaster resilience:** This project develops transdisciplinary modes of thinking and problem solving in the areas of disaster and risk management. It brings together faculty and students in civil and environmental engineering, urban planning, information technology, and other disciplines.
5. **Coastal hazards:** Led by professors in environmental engineering and geosciences, this project examines physical features and mitigation techniques in coastal areas of the US.
6. **Health and social issues in rural communities:** This project investigates patterns of change in small rural communities in the United States from 1980 to 2010 to understand why these communities pose challenges for policies and programs in public health and economic and social well being. The team includes scholars from sociology, history, economics, and geography.
7. **Urban green infrastructures:** The team is developing geographic information systems (GIS) layers for biophysical and social data associated with forests, food gardens, and other components of green infrastructure, and evaluates environmental risks and socioeconomic vulnerability.
8. **Indiana maps:** A large collection of Indiana GIS data that is the result of collaborative work of partners from government organizations and universities. The purpose of these efforts is to acquire, improve, and deliver a wide variety of GIS data for Indiana.
9. **Understanding environmental histories:** The team focuses on water quality using a “citizen science” approach, engaging residents and community organizations to collect and test water samples supplemented with surveys and interviews to better understand narratives about water and the environment.

Our proposed project will engage with each use case through four iterative stages: assessment, development, implementation, and collaborative evaluation.

Stage 1. Assessment of current practices.

The project will begin with a kick-off meeting with use case project teams at each institution. During these meetings we will introduce our team members, explain the goals of the study, and develop a schedule for data collection, which will include participant observation and interviews.

Participant observation will include identifying appropriate social situations, participating in them, performing observation, taking ethnographic notes, analyzing and theorizing, and communicating results. In addition to meetings as the main and most accessible social situation for participation and observation, we will include other situations in our observation repertoire, for example, situations in which use case team members interact with each other regarding data work (e.g., data collection or analysis events). Digital communications and products, such as data instruments and processing techniques and documentation, will be included as part of the data to be collected.

Semi-structured interviews will be conducted with 10-15 team members from each use case. The purpose of these interviews is to better understand IHCR research and data practices, the associated challenges, and ways to address them. The interviews will draw on the thematic framework (see Figure 1) and will be designed to explore datasets, related cyberinfrastructure, and actors involved in interdisciplinary research. The core questions will cover characteristics, activities, and outcomes associated with the main stages of the research data lifecycle and adapt questions from domain-oriented data curation profiles that aimed at collecting information about the origin and lifecycle of a dataset [18], [19].

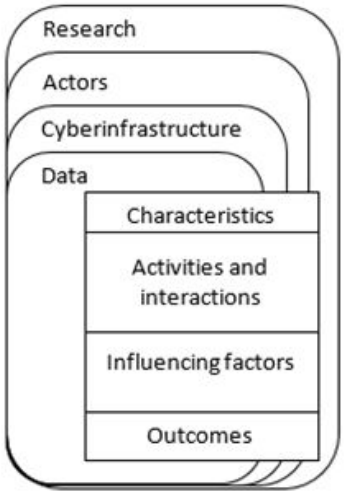

Figure 1. A guiding thematic framework for interviews.

Once the interviews are completed, they will be transcribed, supplemented with short post-interview memos and reflections, and prepared for further analysis. To ensure methodological continuity across all the use cases, the PI will visit each partner campus prior to the interviews, attend several observational situations, discuss data collection procedures, and document and resolve any existing inconsistencies or outstanding questions.

In addition to data collection within and across use case projects, we will collect expert opinions about IHCR data and ways to support it. Expert data collection will follow a combined focus group methodology where groups interact on a topic determined and facilitated by the researcher and where discussions iterate over a series of questions and record the extent of consensus and diversity among the participants [20], [21]. This round of data collection will include both in-person and online panels. The face-to-face panel will convene approximately 10-15 people and will be organized in November 2020 at the Research Data Alliance (RDA) plenary, an international meeting that gathers researchers and professionals interested in data from around the world. Online panels will be convened a month later and will include a larger network of stakeholders, approximately 30-40 experts.

Expert stakeholders will be identified according to the national need discussed above. The following groups will be invited to participate in the panels: data producers, data managers, and data consumers. To avoid geographic and disciplinary biases, the main criterion of expertise will be defined as years of experience with interdisciplinary data (five or more). Additionally, expertise will supersede the criterion of equal representation of each stakeholder group.

To begin panel data collection, experts will be provided with three questions:

- What are the greatest challenges in curating IHCR data?
- What are the barriers to expanding existing curation services to IHCR data?
- What are the most effective strategies for IHCR data curation, including technical tools and human interactions?

The discussions around these questions will be aggregated into common themes and shared back to the experts for revisions. In a short follow-up survey, experts will be asked to review and revise these themes and evaluate rankings of barriers and strategies for IHCR data derived from the prior analysis.

The materials collected during use cases and expert data collection will be stored in a database and analyzed using the grounded theory approach [22]. Data analysis will begin concurrently with data collection to ensure

that all the necessary themes are sufficiently explored and captured. A sample from each data source will be examined independently by two researchers from our team and coded by grouping into similar and related text elements, guided by our research questions. A provisional codebook will be developed based on the analysis of relevant text, repeating ideas, and initial themes. Coding and codebook development will proceed iteratively by discussing thematic overlaps and deviations in interpretations until the researchers converge on a reasonable number of categories. The analysis will then proceed to generate theoretical constructs and hypotheses about common IHCR data practices.

Additionally, the data collected from use case observations and interviews will be used in journey mapping, i.e., in determining user pathways (journeys) in using research technology and library services [23]. Researcher interactions with research technology and library services will be mapped visually and compared to existing pathways and to “ideal” pathways described in the literature. The identified gaps and differences will be used to develop workflows (see Stage 2) and explore changes that would improve researcher - data manager / librarian interactions during the implementation study and beyond.

To ensure rigor and trustworthiness, an audit trail will document all the modifications to protocols, interviews, codebook, themes, and interpretations. Additionally, the researchers will document their perceptions and possible biases and periodically revisit them during the study. At the end of the study, use case teams will have the opportunity to discuss its findings and provide feedback, which will be incorporated into the final interpretation of the data. This step will help to ensure that the findings are consistent with and reflect the participants’ experiences [24].

Stage 2. Workflow development.

Workflow development will start with an inventory of the existing data curation infrastructure within each collaborating institution. As research organizations, these institutions have already built complex processes and cyberinfrastructure (CI) to support research and data curation [25]. We will consult the campuses’ CI plans, data management policies, and guidance on data management plans to develop an inventory of relevant CI resources. Additionally, members of our team will reach out to research technology and library personnel on their campuses to supplement and verify the information we gather.

The inventory work will follow a typical research project lifecycle and include survey of tools, techniques, and services for such categories as data collection, storage, research compliance, content sharing, analytics, communication, collaboration, project management, and dissemination of results. Mapping of both CI components and data-related services is crucial in understanding what campuses are already offering, what can be re-used, and where gaps exist. This information will also be useful in giving researchers a more comprehensive view of the resources available to them at different stages of the research process.

The practices and user journey maps identified in Stage 1 will serve as the foundation for workflow development. For the purposes of this project, we understand a workflow as an intersection between repeatable steps in processes that may or may not be automated and steps that support those processes that can be executed automatically [26], [27]. The user journey maps will be formalized into user requirements, i.e., short descriptions of features that are documented from the user perspective. The descriptions will include information about the process, who performs it, and what information or resources are needed to perform it.

For example, a requirement could describe data analysis, which team members need access to what tools, and how they transfer data to perform such analysis.

The next step in workflow development will be to align user requirements with the existing data curation infrastructure. The alignment will be customized to each campus, but we will also identify common patterns in curation needs and approaches as well as issues to address through additional development or training. The alignment process will also follow the research lifecycle, with particular emphasis on identifying or clarifying curator roles during each stage of the lifecycle [28].

Workflows will leverage existing, well-established tools, such as institutional repositories, scholarly commons, and file storage and sharing. The development will cover automated and “human-in-the-loop” processes and will emphasize data documentation and workflows between collaborators and repositories. We expect a significant amount of work to focus on the development of appropriate training and documentation techniques as one of the bottlenecks in IHCR curation [29].

We will also incorporate in-house, customized tools specifically designed to serve a particular purpose for the host organization. For example, Indiana University is developing a custom metadata management tool that is currently being rolled out to a group on campus. It is particularly suited for interdisciplinary research data as it does not impose specific file formats or disrupt existing processes. This and other tools will be generalized for a greater variety of scenarios and made available for testing across all use cases in this project. Special attention will be paid to cross-pollination between participating institutions.

After the initial round of development and documentation, the workflows will undergo internal testing and validation walk-throughs. The developer on the team will lead other members through the workflows, while the team discusses possible gaps, errors, and other problems. Feedback from these walkthroughs will be used for workflow optimization. Finally, the developed workflows will be documented and prepared for implementation.

Stage 3. Implementation study.

The implementation stage integrates action research [30] and the theoretical framework of designing human practices [31] to a) connect the results of research and development with those who benefit from it, and b) reflect upon the impact implementation activities will have on the participants and their processes. In our case, we will connect our own research and workflow development with the use case projects. The data we collect in Stages 1 and 2 of the project will be used to design the activities of the implementation stage to create change and make space for contributions from all participating projects and their team members, rather than focusing exclusively on creating generalizable knowledge [32], [33].

The implementation study will begin with the design of an implementation blueprint that will describe the study’s purpose, nature, and scope of the change, the degree to which the workflows can be adapted as well as the roles and resources needed for implementation and adaptation. The implementation will take place as a series of iterations that will follow the stages of organizational innovation adoption: awareness, consideration, intention, decision, and continuous use [34].

To initiate implementation through awareness, we will prepare presentation materials and organize educational meetings with each use case team. During these meetings we will demonstrate the workflows and discuss how and to what extent teams can adopt these workflows. Since this is an intervention into existing practices and behavior, the discussion will focus on adaptation and on whether there is a need to identify core components

that are essential to keep, and peripheral components that can be omitted without compromising the usefulness and validity of implementation across all sites [35].

To encourage consideration of workflow adoption and to document intentions and decisions, our team will be available for consultations throughout the duration of the project to discuss the workflows, support implementation, and collect answers to common questions that arise. The feedback will be used to further modify and generalize the workflows. During the implementation stage we will continue participant observations and collect additional data about each implementation setting, participants (including researchers and assisting professionals), resources, team culture, and interactions during implementation. Feedback about the progress and quality of implementation will be collected along with regular individual and team debriefing about progress and experience. The outcomes to be documented will include fit for use, feasibility, likelihood of adoption, implementation cost, scalability (measured as willingness to recommend the workflows to other teams), and sustainability. The last stage of the adoption framework, continuous use, will be approximated through questions such as whether the team is interested in incorporating the proposed workflows into their regular activities.

Studies of implementations and innovation adoption recognize the complexities of any behavior change and the risks of failure or reduced impact of implementations [36]. While our project aims to initiate change at the use case team level, working with individuals who have different attitudes and responses to behavioral change may add to these risks. To mitigate these risks, we will rely on open and transparent communication and dissemination as well as on ongoing support and reflection. Engaging with use case teams early on and developing trust in data managers and curators will also help with improving the outcomes of implementation [37].

The workflows will not be simply handed to the use case teams. As we anticipate both human and technical challenges, we will interact with use case team members and address them as they arise. Attempts to implement the workflows will be considered the beginning of an ongoing relationship with the use case teams that will be nurtured by our own team members and, subsequently, by our institutions' library and data professionals. We will also document challenges and mitigation techniques that will later be published or become part of our best practices reports.

Once the modified workflows are documented, we will develop guidelines for their use and prepare them for collaborative evaluation.

Stage 4. Collaborative evaluation.

Evaluation activities provide a systematic way to understand a project's processes and outcomes [38]. We will focus evaluation activities on project completion, assessing its potential for long-term impact and scalability, identifying outcomes strengths and weaknesses, and seeking feedback on improvement. This is a specialized form of research that integrates well with our case study and action research approaches, as it gathers data to evaluate the project rather than creates generalizable knowledge. Our approach to collaborative evaluation, however, differs from the approaches often discussed in the literature, as it broadens the notion of stakeholders in the project and engages potential and future service providers in the evaluation process [39].

This stage will begin with a thorough review of the project and preparation of documentation and guidelines. We will write up the best practices of IHCR data curation, publish the workflows and associated training materials, and disseminate project study materials for student and peer evaluation.

For student testing and evaluation we will engage graduate students from LIS schools and data science programs in the US to review and test the curation workflows and suggested practice modifications. We have been working with faculty instructors at Indiana University and University of Michigan (see letters of collaboration) and will reach out to more schools that have strong data curation programs, including the University of Illinois at Urbana Champaign and the University of North Carolina. The feedback generated by students will contribute to the further adaptation of workflows and improvement of documentation. In addition to contributing to the project, the students, who are the future data curators and librarians, will gain experience working with challenging real-world examples in data curation.

For peer review and feedback, the workflows and best practices will be disseminated within several professional networks. First, we will collaborate with the Data Curation Network (DCN), a network of ten academic institutions that share diverse expertise in data curation and technologies. We will also collaborate with the Research Data Alliance (RDA) and the Council on Library and Information Resources (CLIR) to disseminate our results and promote peer evaluation. Librarians from the Carnegie Mellon University Libraries and participants of the Ostrom Workshop at Indiana University have also agreed to work with us and help in dissemination and evaluation (see letters of collaboration). Additionally, the results of our work will be presented at professional conferences, including the Research Data Access & Preservation Association (RDAP) summit, International Association for Social Science Information Services and Technology (IASSIST) conference, and other events. One of these presentations will be organized as a special peer evaluation panel with invitations to experts whose feedback is especially valuable, including researchers and data professionals involved in IHCR.

To engage broader communities, we will also organize a virtual workshop that will provide training in interdisciplinary data curation and invite interested libraries to evaluate the model at their institutions. Several libraries have already expressed interest in participating, and we will make a special effort to reach out to and involve smaller academic institutions. The use of these knowledge and professional networks will improve our overall approach and recommendations, adapt it to multiple contexts, and strengthen the national impact of the project.

At the end of the project, we will prepare final documentation on workflows and data curation best practices as well as manuscripts for publication. In addition to publishing in and presenting at traditional academic venues, we will use alternative venues, such as social media, library websites, and discussion forums.

Diversity Plan

The proposed project serves both researchers and the broader data management community, which includes library professionals, data scientists, and IT professionals. Our selection of use case teams emphasizes the diversity of demographic characteristics, such as race, gender and ethnicity, institutional backgrounds, and research stages of the projects. We recognize that research is a complex activity that is performed not only by tenure-track faculty, but also by research scientists, students, and professional staff. Both our own team and

the use case teams consist of researchers and staff of various demographic backgrounds and will include senior, mid-level, and junior personnel. These criteria are built into this proposal.

The projects selected as use cases similarly represent diversity in human experience and interest. The use cases were identified to cover a broad range of issues, with particular emphasis on those relevant to vulnerable or underrepresented groups, including such issues as community resilience, portrayal of enslaved peoples, or environmental health. In the dissemination phase, we will make a particular effort to highlight findings related to projects and data of relevance to these groups.

Both the research and data management communities will be engaged in this project. As it combines qualitative research with design, evaluation, and a broad outreach through expert networks and events, many members of these communities will be involved in defining the needs and opportunities for curating IHCR data. Our use cases will also be useful in data curation training and provide an empirical foundation for such training. For testing and peer evaluation, we will recruit graduate students, postdocs and early professionals of diverse backgrounds, contributing to the transformation of career pathways and enhancing workforce diversity in data curation. Broad outreach and peer feedback will also help in tracking the progress of the project.

Our plan to ensure diversity:

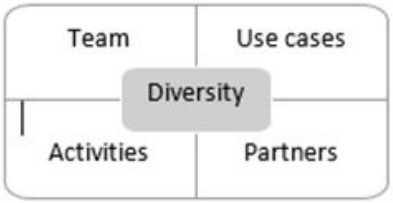

Figure 2. Approach to diversity.

- Work with diverse projects as our use cases
- Engage wide demographics in collaborating research institutions
- Incorporate diversity and inclusion as themes in our research and evaluation activities
- Engage professional networks for dissemination and evaluation of findings and outcomes
- Disseminate findings to smaller institutions that may have fewer resources but encourage faculty to collaborate and engage in scholarly activities
- Engage diverse students in library science, informatics, data science, and similar domains and provide mentorship and learning opportunities through project activities
- Create an inclusive and intellectually diverse working environment
- Consider alternative forms of collaboration and dissemination, including social media, virtual environments, and open platforms

National Impact

The project pursues a two-pronged goal. On one hand, it aims to understand how interdisciplinary researchers work with data and what can be done to help them improve the quality and usability of IHCR data. On the other, the project will create a suite of best practices and workflows in support of data curation that are generalizable across libraries and will empower librarians to become partners in IHCR. This, in turn, will increase library capacity to support heterogeneous data and improve inter-institutional communication and collaboration. The information gathered will be used to create a generalizable data curation model that harnesses existing university services, infrastructure, and engagement mechanisms to support data curation.

The project's **national impact** will be demonstrated in: a) increased knowledge of IHCR data practices that will allow libraries to expand data services and promote best practices in data curation; b) practical guidance in developing services and connecting them to user needs; and c) new collaborations between libraries,

researchers, and data management professionals. The focus on workforce training and collaborative evaluation will result in greater impact than an individual library could achieve. Ultimately, the insights gained from this project will contribute to the maturing of data curation and sharing cultures, to the increased impact of IHCR, and to systemic changes within IHCR and data management communities.

Project deliverables include publications (papers, presentations, and promotional materials for future engagement) that synthesize findings from our studies of IHCR data practices, code and documentation of workflows, and white papers that describe best practices in IHCR curation. Additionally, we will provide recommendations that will enable new or existing curation services to scale their operations to meet interdisciplinary research needs. Moreover, we will develop data management plan templates for interdisciplinary research and training materials for curators and researchers interested in IHCR.

To ensure that our deliverables are adaptable, we will share them at earlier stages of the project and encourage experimentation within use cases and as part of student projects. We will also do the following to facilitate adoption and flexibility of the outcomes: a) provide thorough documentation, b) organize training webinars and make them available long-term through our institutional repositories, and c) incorporate scenarios into our documentation that address the divergences among the examined use cases and propose paths for further adaptation. We will also offer consultations to researchers and data managers throughout the project’s duration and limited consultations after the project is completed.

Our training and dissemination activities will maximize the adoption of workflows and best practices recommendations by emphasizing that incremental rather than radical changes are needed to improve the practices of specific IHCR teams. Dissemination among library and data management communities and discussions of how to implement IHCR curation as a service will promote connections between researchers and librarians and facilitate mutual learning and trust. The documentation and outputs will also allow others to understand how they can replicate our work.

When the funding period ends, components of our workflows and best practices will be incorporated into the library services of collaborating institutions. They will be described through help pages and LibGuides and will be included as links in data management plan templates. We will also make our institutions' data management teams aware of our work and encourage them to incorporate it into consultations with researchers as appropriate. As many members of our team are involved in campus-wide efforts to support researchers in their data curation efforts, our findings will contribute to campus conversations on the highest level.

Additionally, IU Research Technologies will maintain a dedicated web page where the project deliverables will be aggregated and described. Publishing the findings in journals and conference proceedings, especially through open access platforms, will ensure that they are available for the wider community and can be used in further studies of IHCR data practices and workflows. Working with DCN, CLIR, and the RDA Libraries for Research Data (L4RD) to make our recommendations part of official recommendations and documentation will leverage these diverse networks for further adoption of our deliverables.

SCHEDULE OF COMPLETION

	Year 1 (Sep 2020 - Aug 2021)				Year 2 (Sep 2021 - Aug 2022)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1. Assessment of current practices.								
Kick-off discussion with each team	X							
Joint project meetings and discussions	X	X	X					
Data practices interviews		X	X					
Campus visits by PD to collaborating institutions	X	X						
Expert panels	X		X					
Data analysis and user journey mapping			X	X				
Common practices documentation				X				
Dissemination (papers, presentations)					X	X	X	
2. Workflow development.								
Inventory of existing data curation infrastructure	X							
Development of user requirements		X	X					
Alignment between requirements and existing data curation infrastructure			X					
Design and development of data lifecycle curation workflows			X	X				
Internal testing (through discussions and walkthroughs)				X	X			
Feedback analysis and workflow optimization					X			
Workflow documentation				X	X	X		
3. Implementation study.								
Design of implementation blueprint			X	X				
Workflows demonstration to each team				X	X			
Implementation in use case teams					X	X		
Consultations and feedback					X	X	X	X
Workflows adaptation							X	
Documentation and guidelines updates							X	X
4. Collaborative evaluation and training.								
Best practices review and documentation				X	X	X	X	
Student testing					X			
Peer reviews and feedback (DCN, RDA, CLIR and academic libraries)						X	X	
Training workshops and webinars							X	X
Peer evaluation event						X	X	
Final documentation on workflows and best practices in IHCR data curation							X	
Dissemination (papers, presentations, code, documentation)								X

DIGITAL PRODUCT FORM

INTRODUCTION

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to digital products that are created using federal funds. This includes (1) digitized and born-digital content, resources, or assets; (2) software; and (3) research data (see below for more specific examples). Excluded are preliminary analyses, drafts of papers, plans for future research, peer-review assessments, and communications with colleagues.

The digital products you create with IMLS funding require effective stewardship to protect and enhance their value, and they should be freely and readily available for use and reuse 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

If you propose to create digital products in the course of your IMLS-funded project, you must first provide answers to the questions in **SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS.** Then consider which of the following types of digital products you will create in your project, and complete each section of the form that is applicable.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

Complete this section if your project will create digital content, resources, or assets. These include both digitized and born-digital products created by individuals, project teams, or through community gatherings during your project. Examples include, but are not limited to, still images, audio files, moving images, microfilm, object inventories, object catalogs, artworks, books, posters, curricula, field books, maps, notebooks, scientific labels, metadata schema, charts, tables, drawings, workflows, and teacher toolkits. Your project may involve making these materials available through public or access-controlled websites, kiosks, or live or recorded programs.

SECTION III: SOFTWARE

Complete this section if your project will create software, including any source code, algorithms, applications, and digital tools plus the accompanying documentation created by you during your project.

SECTION IV: RESEARCH DATA

Complete this section if your project will create research data, including recorded factual information and supporting documentation, commonly accepted as relevant to validating research findings and to supporting scholarly publications.

SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

A.1 We expect applicants seeking federal funds for developing or creating digital products to release these files under open-source licenses to maximize access and promote reuse. What will be the intellectual property status of the digital products (i.e., digital content, resources, or assets; software; research data) you intend to create? What ownership rights will your organization assert over the files you intend to create, and what conditions will you impose on their access and use? Who will hold the copyright(s)? Explain and justify your licensing selections. Identify and explain the license under which you will release the files (e.g., a non-restrictive license such as BSD, GNU, MIT, Creative Commons licenses; RightsStatements.org statements). 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.

The formal products produced during our projects are research studies data and documentation, training materials, computer code, and user and developer documentation. We also anticipate intermediate products emerging as a result of conducting our work and documenting our progress. The formal materials and software products resulting from this effort will be licensed using open and free licensing, e.g., Creative Commons and Apache 2.0-style licenses. Intermediate products will be discarded by the end of the project life or re-used to become part of the formal products. We will adhere to FAIR guidelines in maximizing access to our formal products.

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.

Our team will work with Indiana University data stewards to ensure that computer code developed in this project can be shared openly without violating any confidentiality or sensitivity policies (for example, if the workflows enable storage and transfer of restricted or embargoed data). The code and associated training materials will be shared broadly via open software repositories and IU sharing platforms.

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.

As part of this project, we will be conducting interviews and taking notes during ethnographic observations. The data collected via interactions with human subjects will be stored securely and accessed by project investigators only. Such data will be shared only after appropriate de-identification and with explicit consent from participants.

SECTION II: 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.

In the course of this project the following digital content will be created:

1. Workflow components and modifications - computer code.
2. Workflow components and modifications - human instructions.
3. Interview recordings and transcripts and field notes. (See Part IV Datasets for more details.)
4. Online manuals and education and training materials.

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

This project does not focus on creating digital content beyond computer code, research studies data and documentation and training materials (presentations and reports). We will use computers and software at Indiana University and partner institutions.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG, OBJ, DOC, PDF) you plan to use. If digitizing content, describe the quality standards (e.g., resolution, sampling rate, pixel dimensions) you will use for the files you will create.

Computer code will exist in software formats, such as Java, Javascript, Python, R, XML and HTML. Observational data and interviews will be recorded in MPEG and text formats (DOC, TXT). Quality standards will be ensured via iterative testing of computer code and via regular checking procedures for other formats.

Workflow and Asset Maintenance/Preservation

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

The project is led by Dr. Kouper who has a strong record of publishing peer-reviewed research in relevant areas. Processes and products will be regularly evaluated for completeness and consistency, ensuring that recordings and other data are properly documented and saved. Software development will be supervised by co-PI Tuna and code review sessions will be organized as needed. The team will meet periodically to discuss quality issues and ensure that the work and products are aligned with the proposed scope. An additional quality check will be performed regularly during meetings with use case teams.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period. Your plan should 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).

Project digital assets will be maintained using Indiana University storage systems (Google drive, OneDrive and Microsoft Teams, GitHub). To preserve and share the formal products of the project, we will use institutional repositories of each campus. Products of research (publications, data, presentations) will also be shared through traditional publication venues.

Metadata

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

To ensure interoperability and ease of use, the project will rely on REAME files structured to include main metadata information (dates, authorship, provenance, scope, units and so on). Computer code will also be supplied with documentation and software metadata.

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

Metadata will be maintained as part of the digital asset maintenance, embedded where possible and linked where embedding is not possible. Metadata will be considered part of the digital assets and research objects and migrated together as needed.

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).

As the project is not concerned with creating digital collections and platforms, we will rely on existing channels for widespread dissemination and use: partner organizations and their websites, open publication venues, conferences, listservs and institutional repositories.

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, delivery enabled by IIIF specifications).

Computer code and study products will be openly available online, unless restricted by the publishing entities or limited by the requests of confidentiality.

D.2. Provide the name(s) and URL(s) (Universal Resource Locator), DOI (Digital Object Identifier), or other persistent identifier for any examples of previous digital content, resources, or assets your organization has created.

Most of the work published by Kouper and others has been published with persistent DOIs (see resumes for examples).

SECTION III: SOFTWARE

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.

The software products will include a set of executable workflow descriptions and meta data annotations, and tools to generate, derive and transform these descriptions and metadata. The workflow descriptions will be written in Python. We will leverage existing open source software for the generation, execution and analysis of the workflows and metadata.

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

There are many tools and software frameworks, open source or otherwise, that support managing and execution of workflows. However, these systems are not adapted to handle workflows that support metadata curation and management. As our workflow development focuses on interdisciplinary data curation, focus on metadata and workflow customization are important features and this is a significant difference between our work and other software.

Technical Information

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

Python will be the primary workflow development language. These workflows will be scheduled and executed using the Apache Airflow. RabbitMQ will be used to handle messaging between processes that span across multiple processing and storage systems. MySQL will be used to store both user project metadata as well as operational system data. Apache Solr will be used for search features. Angular.JS will be used to create simple yet functional user interfaces to various system features.

Some of these platforms and frameworks may be adapted based on the requirements of each partner institution.

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

Executable code will be in Python, that is consumable by most execution engines and schedulers. Data import and export will be in XML and JSON formats. Metadata will be structured in RDF where appropriate (in addition to README files described above that are more appropriate for human consumption).

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

As project data will be published in institutional and software repositories, there is a dependency on those platforms and their APIs.

Additionally, software execution may depend on admin access rights configuration, which will be verified during the testing and implementation stage.

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

The development will be managed in IU's institutional Atlassian/Jira using Kanban boards. All associated code and documentation will be managed via a Github repository.

B.5 Provide the name(s), URL(s), and/or code repository locations for examples of any previous software your organization has created.

The Collectome software is the most recent grant-funded software system development project that co-PI Esen Tuna was involved in. The details of the project and the associated open source code developed can be found here:
https://collectome.rt.iu.edu/frontend/home
https://github.com/dmreagan/collectome/.

Access and Use

C.1 Describe how you will make the software and source code available to the public and/or its intended users.

The software, source code and associated documentation will be made available on github.com project in Indiana University's public repository.

C.2 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

Indiana University GitHub

URL:

https://github.com/indiana-university

SECTION IV: RESEARCH DATA

As part of the federal government's commitment to increase access to federally funded research data, Section IV represents the Data Management Plan (DMP) for research proposals and should reflect data management, dissemination, and preservation best practices in the applicant's area of research appropriate to the data that the project will generate.

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

Data will be collected via observations and interviews, which includes note-taking, recordings, videos, and photographs. The data will be analyzed using grounded theory approach, qualitative coding, and content analysis. Data will be collected and analyzed iteratively throughout the duration of the project (e.g., during stages 1, 3 and 4).

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?

Data collection involves human subjects and requires IRB approval. IRB application will be prepared and submitted when / if the project is approved for funding.

A.3 Will you collect any sensitive information? This may include personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information. If so, detail the specific steps you will take to protect the information while you prepare it for public release (e.g., anonymizing individual identifiers, data aggregation). If the data will not be released publicly, explain why the data cannot be shared due to the protection of privacy, confidentiality, security, intellectual property, and other rights or requirements.

Project participants (use case team members) can be identified during data collection. Personally identifiable information will be stored securely and only the project director and key personnel will have access to it. Before public release of the data all PII will be removed (participants will be assigned coded numbers and any information that may identify them individually will be obscured in the interviews, notes, and transcripts).

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding retrieving, displaying, processing, or otherwise reusing the data?

Data will be shared in open text-based formats so that any software can be used to analyze it. Computer code may require software to run it.

A.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

Participants will be provided with study information sheets. Codebooks and data notes will be created as part of the analysis of qualitative data (e.g., in the thematic coding procedures). Codes, their descriptions and other documentation that describes when and where the interviews and observations took place will be stored in text formats along with the data. The documentation will be associated with the datasets through consistent file naming and through identifiers that refer to each data collection effort separately.

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

The data will be managed and archived using Scholarly Data Archive (backed-up storage for long-term archiving) and institutional Google Drive at Indiana University (for active work with data). Folders with appropriate permissions for data, processing scripts, IRB documentation, and publications will be created. For dissemination, we will use institutional repositories at each institution.

A.7 Identify where you will deposit the data:

Name of repository:

IU Scholarworks
CU Scholar

URL:

scholarworks.iu.edu/dspace
scholar.colorado.edu

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

The plan will be reviewed and monitored by project director and co-directors. We will review it every 6 months and adjust according to the amounts and types of data generated.