Abstract

The Montana State University (MSU) Library seeks a $25,000 Sparks Grant in the Curating Collections project category to design, build, and assess a prototype Institutional Research Data Index (IRDI). Many academic libraries now support two repository systems—one for publications, and another for research data—even when there are nearly a thousand data repositories in the United States (see re3data), many of which provide services and policies that ensure their trustworthiness and suitability for institutional research data. We suggest that small and mid-sized institutions can both conserve their limited resources and increase discovery of local research datasets by directing researchers to deposit research data in third-party repositories, and then providing local access to institutional research datasets through IRDI. The IRDI prototype will promote discovery and reuse of institutional datasets while expending far fewer resources than those required for an institutional data repository. IRDI will be an easy-to-install metadata indexing system that (1) automatically harvests metadata from third-party data repositories; (2) generates new descriptive metadata for individual datasets using external topic mining of scholarly profile sources like ORCID and Google Scholar Profiles; and (3) is optimized for discovery by commercial search engines.

The project team is Project Director Sara Mannheimer (Data Librarian), Key Project Staff Jason Clark (Head of Special Collections and Archival Informatics), Key Project Staff James Espeland (Software Engineer), and a student research assistant. During Phase 1 (Foundational activities, Oct. 2018-Nov. 2018), we will advertise and hire a student research assistant position, develop the metadata model, investigate options for automatic metadata harvesting and generation, create a Github repository and project website, and host the first of two teleconferenced advisory board meetings. During Phase 2 (Prototype startup, Dec. 2018-May 2019), we will build the initial IRDI prototype with RSS and API connections, solicit feedback at Code4Lib in Feb. 2019, and conduct usability checks. During Phase 3 (Troubleshooting and cleanup, Jun. 2019-Jul. 2019) we will host the second advisory board meeting, investigate solutions to metadata issues, clean up usability issues, and solve any remaining problems. During Phase 4 (Dissemination of results, Aug. 2019-Sept. 2019), we will post the package installer and instructions to our project website, conduct training webinars, and establish a user forum for user community building.

IRDI supports the IMLS performance goal regarding Content and Collections: “Broaden access and expand use of the Nation’s content and collections.” To measure this performance goal, we will gather Performance Measure Data from the IRDI prototype using web analytics. We will also gather Performance Measure Data from user community behavior and feedback using (1) web analytics of the IRDI project website; and (2) qualitative data collected through the website’s feedback form and a user forum.

Ultimately, IRDI will have three key impacts. (1) Increased discovery and reuse for institutional research data; (2) Increased awareness of research data as scholarly output; and (3) Community-building and efficiencies for small and mid-sized institutions. While the currently-scoped prototype will be disseminated via individual installations, if successful, the prototype could theoretically be expanded into a community-administered index, leading to economies of scale.

A Prototype for an Institutional Research Data Index

Summary

The Montana State University (MSU) Library seeks a $25,000 Sparks Grant in the Curating Collections project category to design, build, and assess a prototype Institutional Research Data Index (IRDI). IRDI will be an easy-to-install metadata indexing system that promotes discovery, reuse, and impact of institutional research datasets by (1) automatically harvesting metadata from third-party data repositories; (2) generating new descriptive metadata for individual datasets using external topic mining of scholarly profile sources like ORCID and Google Scholar Profiles; and (3) by optimizing for discovery by commercial search engines.

Statement of National Need

Academic libraries are increasingly participating in research data publication and preservation. However, out-of-the-box institutional repository (IR) systems like DSpace and Digital Commons are not designed to publish research data. These systems' workflows are tailored to articles, which are published once, in a final state. Data workflows tend to be messier; research data is often published in a preliminary state, then updated with new versions as projects progress (Xie et al., 2016). Additionally, many IR systems lack data-specific features such as file-level description and data-specific metadata. Customizing IR systems to meet the needs of data may prevent upgrading to new software versions, because any customizations must be repeated for the new version. As an example, Dryad Digital Repository has heavily customized DSpace, and as a result, it still uses DSpace 1, a version first released in 2002.[1]

An increasing number of repositories are being designed specifically for research data, but these systems are resource-intensive. Open source systems like Dataverse, CKAN/DKAN, and Fedora/Hydra/Sufia require developer hours and data storage infrastructure. Vendor solutions like Figshare for Institutions and Tind require subscription payments. And the resources required to operate a data repository are expended in addition to those required for existing IRs. Many academic libraries now support two repository systems—one for publications, and another for research data. In order to support research data repositories, libraries are either increasing spending by buying vendor solutions, or replicating work by building and managing individual instances of data repository software. As of 2018, there are nearly a thousand data repositories in the United States (see re3data), many of which provide services and

[1] Dryad's recently-announced partnership with California Digital Library will include the launch of "a new, modern and easier-to-use platform" (Hull, 2018).

policies that ensure their trustworthiness and suitability for institutional research data. We suggest that small and mid-sized institutions can both conserve their limited resources and increase the discovery of local research datasets by directing researchers to one of these third-party repositories, and then providing local access to institutional research datasets through a research data index.

We propose a $25,000 Sparks Grant to build a prototype for an Institutional Research Data Index (IRDI), a scalable and sustainable metadata index that will promote discovery and reuse of institutional datasets while expending far fewer resources than those required for an institutional data repository. Unlike a data repository, IRDI will not archive research datasets themselves. Instead, it will harvest metadata from third-party data repositories that archive research datasets, and serve the metadata via an online interface. To explain further: in the same way that a library catalog does not store actual books, but rather provides metadata so that visitors can find the books, IRDI does not store actual research datasets, but rather provides metadata so that visitors can find the datasets in third-party data repositories.

National Impact

The IRDI pilot will have three key impacts. First, IRDI will allow institutional research data to be published in discipline-specific repositories, while simultaneously being discoverable in an institutional index, thus promoting increased discovery, reuse, and citation of academic research data. An exploratory study conducted by the PD suggests that research data are more likely to be discovered and reused if they are (1) archived in a discipline-specific repository; and (2) indexed in multiple places online (Mannheimer, Sterman, & Borda, 2016). With IRDI, research data can be published in the data repositories that are commonly used within disciplines, while providing search engine optimization to encourage discovery by the broader community.

Second, IRDI will reinforce the idea that research data are a legitimate scholarly product, and help institutions tell the story of how the research data that originates at our institutions impact the scholarly landscape. IRDI will act as an index of institutional research data, showcasing local data regardless of whether they are published in the IR or in a third-party repository. IRDI will provide information to institutional Research Information Management (RIM)/Current Research Information (CRIS) systems; the PD has met with the MSU Office of Planning and Analysis to discuss metadata interoperability. IRDI will also function as a database that institutions can query for tracking and assessment purposes. Using web analytics, IRDI can track site visits, pageviews, inbound referrals, and outbound links. IRDI metadata will be also available through an API, and can be used as a dataset for institutional assessment. We acknowledge the possibility that IRDI may not provide a fully comprehensive index of all of our institutional research datasets. However, IRDI metadata can still be used

answer questions such as: Which data repositories indexed by IRDI are used most by researchers at our institution? How high-quality are the metadata and documentation for the institutional research datasets indexed by IRDI? What are the collaborations between researchers in the IRDI index? These questions can help institutional administrators better understand the scholarly outputs of the institution.

Additionally, a third, much broader impact could theoretically be realized if the IRDI project is successful. While the currently-proposed IRDI prototype is a locally-installed system that can be adopted by other small and mid-sized institutions, there is a possibility that IRDI could someday be expanded into a single, unified index for institutional research datasets that could be adopted community-wide. Such a system would have a single team of system administrators, and individual institutions would curate the automatically-harvested metadata for their own institution's research datasets. A community-wide institutional research data index would take advantage of economies of scale. By working together to build a system that can be administered by the community at large, rather than building individual systems that are replicated at each institution, we can build bigger, better systems that promote one of our key goals as academic libraries—to provide discovery and access for scholarly products.

Project Design

IRDI will build on related projects such as DataMed, NYU Health Sciences Library Data Catalog, Cinergi, OmicsDI, and SHARE, offering three critical innovations: (1) IRDI brings an institutional focus to the automated collection of metadata from external data repositories. Automated metadata collection allows the index to be populated with metadata for institutional datasets with less manual effort from library employees and therefore less resource expenditure from the institution; (2) IRDI will also automatically generate new descriptive metadata for individual datasets using external topic mining of scholarly profile sources like ORCID and Google Scholar Profiles; and (3) IRDI will be optimized for discovery through leading commercial search engines. The project team has experimented with some early prototyping of the IRDI system, including building a preliminary metadata model informed by Schema.org, Data Catalog Vocabulary (DCAT), DataCite, DATS, and Project Open Data metadata schemas.[2] The final prototype will also have customizable metadata fields to support institution-specific metadata. The IRDI prototype is based on a system developed at MSU Library that uses Javascript and PHP scripts to parse RSS feeds and harvest metadata for institutional publications (Sterman & Clark, 2017).

[2] In a nod to the rising importance and timeliness of our project, we note that metadata schemes for data continue to evolve and be developed. For example, the latest DCAT version was introduced on May 8, 2018.

Project Audience and Beneficiaries

The IRDI prototype will be simple enough to be adopted by small and mid-sized institutions with limited resources. As a mid-sized institution ourselves, MSU has encountered the challenges that IRDI seeks to address. Like many academic libraries, we have ambitions beyond our relatively small size and budget—we want to help researchers at our university publish their research data, and we want that published data to be discovered and reused. IRDI also facilitates an interest convergence between the goals of the Library and of our institutional administration: the Library wants to tell the story of how MSU researchers' openly published data can be used to advance knowledge, and our institutional administration is interested in capturing and measuring the scholarly output for the institution. In addition to advancing the goals of librarians, IRDI will also provide metrics on institutional research data output that can be used to advance the goals of institutional administration. The ultimate aim of IRDI is to be an easy-to-implement system that provides a realistic solution for small and mid-sized institutions looking to fulfill lofty data discovery and assessment goals.

Goals and Outcomes

Goal 1. Create a central index that provides discovery and access for institutional research datasets

- **Outcome 1.** Automated harvesting of institutional research data metadata from data repositories
- **Outcome 2.** Automated generation of metadata for individual datasets using external topic mining of scholarly profile sources
- **Outcome 3.** Optimization of IRDI content for discovery by commercial search engines
- **Outcome 4.** Tracking of institutional research data based on IRDI metadata, including interoperability with RIM/CRIS systems and metrics from third-party data repositories (such as number of views and downloads).
- **Outcome 5.** Development of strategic connections with related projects, including DataMed, NYU Health Sciences Library Data Catalog, and SHARE.

Goal 2. Encourage use of IRDI by our fellow small and mid-sized institutions

- **Outcome 1.** Github repository and package installer for easy installation
- **Outcome 2.** Website for promotion, centralized information dissemination, and community feedback
- **Outcome 3.** IRDI user forum for stakeholder input, consensus building, community building, troubleshooting, feedback, and other communication.
- **Outcome 4.** Training webinars

Long-term goal (beyond the scope of this $25,000 grant). Pursue broader adoption by the academic library community at large, including communal administration of the IRDI system. (For further investigation of this idea, please see National Impact.)

Relevance to IMLS Strategic Goals, Performance Goals, Project Category and Funding Category

The IRDI prototype facilitates discovery and access for research datasets. The IRDI project therefore aligns directly with the IMLS agency-level Goal 3: Increase Public Access (see the IMLS Strategic Plan 2018-2022), especially the objective "invest in tools, technology, and training that enable people of all backgrounds and abilities to discover and use museum and library collections and resources." (For more information on accessibility of the IRDI system and supporting materials, please see Diversity Plan, below.) In addition, the improved discovery and increased access to research datasets that will result from IRDI's development supports IMLS performance goal regarding Content and Collections: "Broaden access and expand use of the Nation's content and collections." Increased discovery and access will lead to increased use of institutional research datasets, thus furthering knowledge and advancing science.

The IRDI prototype also aligns with the Curating Collections project category in its aim to provide increased access to research data, and to encourage use and reuse of research data. If successful, the prototype can scale up to contribute to a National Digital Platform, becoming a regional or national shared service that provides access to research data for the research, educational, and public communities. In so doing, IRDI would support increased efficiencies, cost savings, access, and services, and would create alliances and networks of libraries surrounding discovery of institutional research data. (For further investigation of this idea, please see National Impact).

Sequence of Activities

This work will be completed over the course of one year. Please see the timeline below for the sequence of activities.

Project Timeline

Phase	Duration	Activities
Phase 1. Foundational activities	October - November 2018	• Advertise and hire student position • Develop metadata model • Investigate options for automatic metadata harvesting and generation • Create Github repository and project website • Advisory Board meeting 1

Phase 2. Prototype startup	December 2018 - May 2019	• Build initial prototype • Solicit feedback at Code4Lib, Feb. 2019 • Conduct usability checks
Phase 3. Troubleshooting and cleanup	June 2019 - July 2019	• Advisory board meeting 2 • Investigate solutions to metadata issues • Clean up usability issues • Solve any remaining problems
Phase 4. Dissemination of results	August 2019 - September 2019	• Post package installer and instructions to project website • Conduct training webinars • Establish user forum

Time, Personnel, and Financial Resources Needed to Complete the Project

Our team will consist of three MSU Library employees and a student research assistant.

Project Director (PD) Sara Mannheimer, Data Librarian. Mannheimer has expertise in data repository development and data discovery, and will act as project manager. 4% of Mannheimer's time will be dedicated to the IRDI project.

Key project staff Jason Clark, Head of Special Collections & Archival Informatics. Clark has expertise in metadata and semantic web, and will act as metadata and semantic web lead. 2% of Clark's time will be dedicated to the IRDI project.

Key project staff Jim Espeland, Software Engineer. Espeland has expertise developing library applications and systems, and will act as lead developer. 12.5% of Espeland's time will be dedicated to the IRDI project.

Student research assistant. A junior or senior undergraduate student majoring in computer science will be recruited to help with IRDI development, under Espeland's mentorship.

Advisory Board and Key Partnerships

We have reached out to related data discovery projects DataMed, NYU Health Sciences Library Data Catalog, and SHARE to act as an advisory board for the project. The advisory board will have two teleconference meetings during the duration of the project. We have secured a commitment from Jeff Grethe of DataMed and Kevin Read and Nicole Contaxis of NYU Health Sciences Library. Read and Contaxis have also invited us to connect with their Data Catalog Collaboration Project. Additionally, we have established a partnership with the MSU Office of Planning and Analysis—which manages the MSU RIM/CRIS system—to ensure that IRDI is interoperable with

RIM/CRIS systems.[3] Throughout the project, we will be in communication with the data curation community. We will also be in communication with other data curation-related projects such as the Sloan-funded Data Curation Network and personnel from data repositories—the PD has existing collaborative relationships with ICPSR and Qualitative Data Repository (Mannheimer et al., 2017), and with Dryad Digital Repository (Mannheimer & Hull, 2017).

Stakeholder Input, Consensus Building, and Community Building

The project will be presented at Code4Lib 2019, a key conference for library developers. Code4Lib will take place partway through the project timeline—in February 2019. The Code4Lib presentation is an opportunity for the project team to solicit stakeholder input, build consensus, and create community buy-in around the in-progress IRDI prototype, ultimately leading to a stronger prototype. The IRDI project website will also include a contact form and a link to a user forum for IRDI users. The contact form where users can receive individual assistance or provide feedback on the system, and the user forum will facilitate stakeholder input, consensus building, and troubleshooting for the IRDI system. Community building via Code4Lib and user feedback will support the sustainability of IRDI beyond the conclusion of the funding period.

Dissemination of Project Deliverables to the Project Audience

We will create a project website with links and information about IRDI, installation documentation, and FAQs. IRDI code will be given a Creative Commons Attribution (CC BY) license and released via Github, and IRDI metadata will be available openly via RSS feed and an API. A package installer will be provided so that small and mid-sized institutions with limited developer expertise can easily install the system. The project team will also host training webinars to guide users through installation and implementation of the IRDI prototype.

Risks, Assumptions, and Risk Mitigation

Risk and assumption 1. Data repositories often do not require disclosure of institutional affiliation. Without institutional affiliation in data repository metadata, it is difficult to automatically harvest institution-specific content, and the records in IRDI may not show a complete picture of all of the institutional research datasets that have been published. This project assumes that our team will be able to find and automatically harvest institution-specific research datasets.

Mitigating factors 1. We have identified a few mitigating factors for this risk. First, we have investigated querying data repositories for the full names of MSU researchers.

[3] See attached letters of support from Grethe, Read & Contaxis, and the MSU Office of Planning and Analysis.

Because full-name query is labor-intensive and fails to account for name disambiguation, we have also investigated ORCID integration. ORCID usage is increasing, and could provide a partial solution; the data repository Zenodo is one notable adopter of ORCID integration. Second, many data repositories provide a metadata field for grant numbers; these grant numbers can be cross-referenced with institutional grant award information to discover datasets that support projects with principal investigators from our institution. Third, Dryad Digital Repository has created a script that cross-references Dryad datasets with affiliated published articles to produce institutional affiliation information; this strategy could be effective for datasets with affiliated published articles. Lastly, we will use dataset metadata services such as DataMed, Cinergi, OmicsDI, and DataCite to more efficiently harvest third-party data repository metadata, and we have reached out to SHARE to learn from their experience aggregating institutional repository metadata.

Risk and assumption 2. Institutional data repositories are often used to archive institutional research data that does not neatly fit into disciplinary data repositories. Since IRDI is a metadata index—not a data repository—it is not designed to store local datasets. This project assumes that we can find a solution for these "outlier" datasets.
Mitigating factor 2. Third-party data repositories exist with broad enough collecting policies to support a wide range of submissions from institutional researchers and students, and some third-party repositories can accommodate even very large datasets. The resources saved by foregoing building a local data repository could allow institutions to subsidize any potential cost to archive unique research data in such repositories.

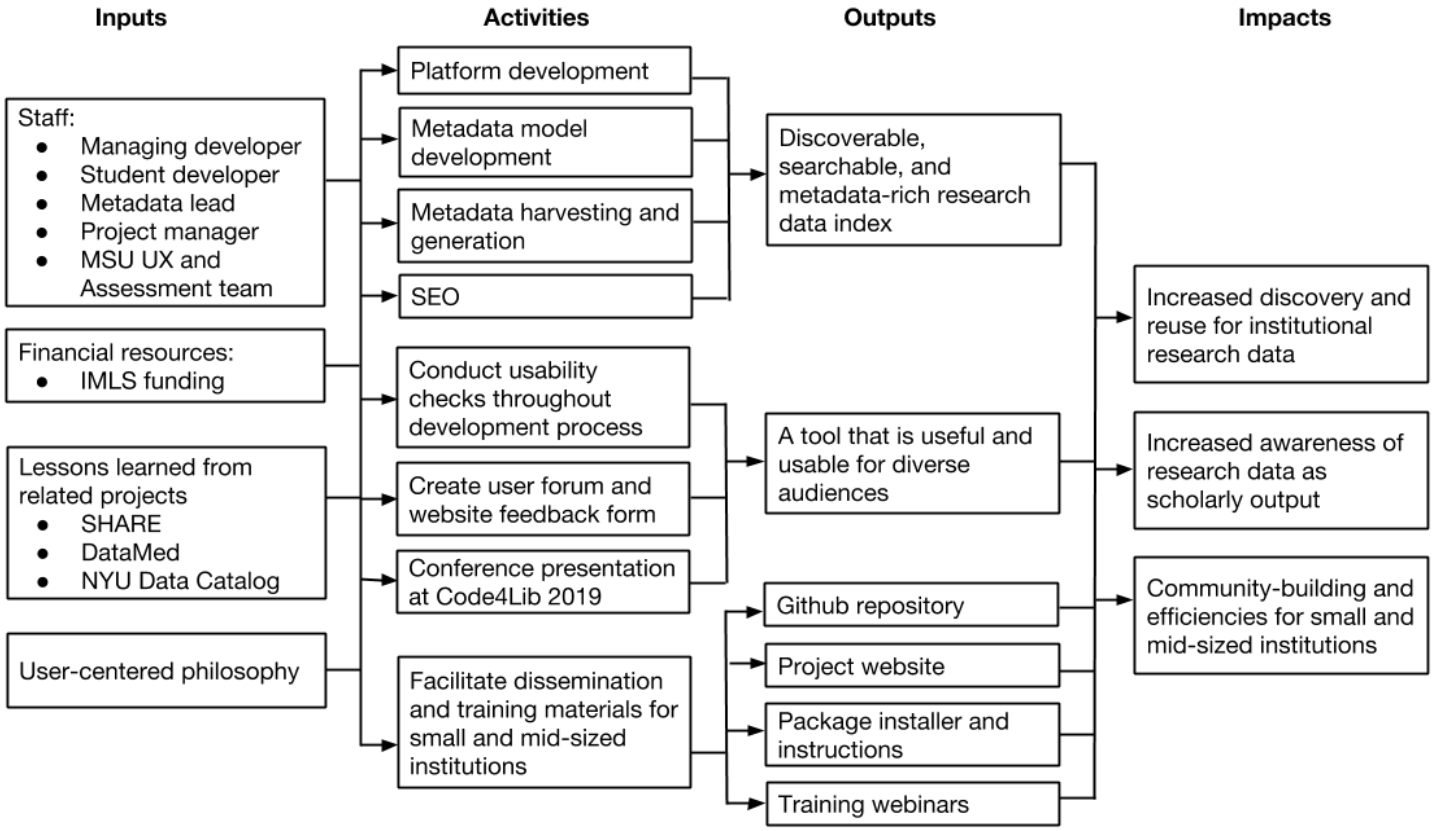
Risk and assumption 3. Search engines such as Google Scholar may not index metadata-only repositories. This project assumes that we will be able to tune the IRDI prototype for indexing by commercial search engines.
Mitigating factor 3. One potential solution is to provide an object in IRDI in order to improve search engine optimization. For example, IRDI could include a downloadable text document containing the dataset metadata and a link to the third-party data repository.

Evaluation and Performance Measurement

Evaluation Logic Model

The logic model below illustrates the inputs, activities, outputs, and impact of the IRDI project. The model will be used to evaluate the success of the project. For a successful project, inputs will lead to activities, which will lead to outputs, which will lead to impacts.

Evaluation Logic Model



Benchmarks for Project's Performance and How They Will be Measured

As illustrated in the evaluation logic model above, success for IRDI hinges on three key outputs: (1) a successful data index prototype; (2) a prototype that is useful and usable for diverse audiences; and (3) the foundations for disseminating the prototype to our community of fellow small and mid-sized institutions. A successful prototype will be measured by the inclusion of the following elements: metadata model that facilitates searching within the index, successful metadata harvesting, and successful search engine optimization so that IRDI content can be discovered through commercial search engines. A prototype that is useful and usable for diverse audiences will be measured by a usable and pleasant user interface and user feedback via our web feedback form and user forum. The foundations for disseminating the prototype to our community will be measured by the creation and use of our project website, Github repository, training webinar attendance, and conference presentation response.

Performance Goal and Performance Measure Data Collection and Analysis

IRDI supports the IMLS performance goal regarding Content and Collections: "Broaden access and expand use of the Nation's content and collections." To measure this performance goal, we will use two key data sources.

- **Performance Measure Data from IRDI prototype.** Once the prototype goes live, web analytics will be used to track IRDI site visits, pageviews, inbound referrals, and outbound links. To protect user privacy, IRDI will activate all available privacy protection features of the web analytics tool (for example, IP anonymization).
- **Performance Measure Data from user community behavior and feedback.** Web analytics will be used to measure visitors to the IRDI project website. Qualitative data will be collected through the website's feedback form, and through the user forum.

Diversity Plan

The IRDI project aims to achieve diversity and inclusion within the scope of our pilot. We understand that diverse perspectives and ideas within the project team will produce stronger deliverables. To help ensure diverse representation and a culture of inclusion within the project team, we will consult with members of MSU's Cultural Attunement Committee, which was created by ADVANCE Project TRACS, a multi-year grant project funded by the National Science Foundation that aims to increase faculty diversity through research, policy, and training. The Cultural Attunement Committee has expertise in recruiting and retaining diverse talent, and the IRDI team will seek their guidance on finalizing the student position description, advertising the job in appropriate forums, and evaluating the student applicants.

The deliverables of our grant—including the IRDI package installer, IRDI prototype user interface, and training materials—will be accessible to those of all abilities and will aim to be inclusive of diverse audiences. To ensure accessibility, we will use assessment tools such as the WAVE web accessibility evaluation tool and the aXe Accessibility Engine to produce accessible content. To ensure that IRDI is inclusive of diverse audiences, we will solicit user feedback to improve our deliverables after they have been disseminated. Feedback mechanisms include an IRDI user forum and a feedback form on the IRDI website. While user research is beyond the scope of the current funding period, if the IRDI prototype is successful, we plan to conduct user research sessions with a diverse range potential users, to ensure that IRDI is useful and usable to the entire scholarly community.

Phase:	Activity:	2018			2019								
		Oct	Nov	Dec	Jan	Feb	Mar	April	May	Jun	Jul	Aug	Sep
Phase 1: Foundational Activities	Advertise and Hire Student Research Assistant Position	■											
	Develop Metadata Model	■	■										
	Investigate Options for Automatic Metadata Harvesting and Generation	■	■										
	Create Github Repository and Project Website	■											
	Advisory Board Meeting 1		■										
Phase 2: Prototype Startup	Build Initial Prototype with RSS and API Connections		■	■	■	■	■	■	■				
	Solicit Feedback At Code4Lib, Feb. 2019					■							
	Conduct Usability Checks				■	■	■	■	■				
Phase 3: Troubleshooting and Cleanup	Advisory Board Meeting 2									■			
	Investigate Solutions To Metadata Issues									■	■		
	Clean Up Usability Issues									■	■		
	Solve Any Remaining Problems									■	■		
Phase 4: Dissemination of Results	Post Package Installer and Instructions To Project Website											■	
	Conduct Training Webinars											■	■
	Establish User forum											■	■

DIGITAL PRODUCT FORM

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

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

PART I: Intellectual Property Rights and Permissions

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

The IRDI prototype code will be assigned a GNU General Public License. All other digital products will be assigned Creative Commons Attribution (CC BY) licenses. GNU GPL and CC BY are minimally restrictive licenses, allowing for maximum reuse, remixing, and redistribution.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The products created under this grant will fall under the ownership of the project team. For more information, please see MSU Faculty Handbook Intellectual Property Policy http://www.montana.edu/policy/faculty_handbook/intellectual_property.html and Montana University System Board of Regents Copyright Policy http://www.montana.edu/policy/faculty_handbook/fh900.html#910.00

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.

No products from this project raise substantial privacy concerns. Once the prototype goes live, web analytics will be used to track IRDI site visits, pageviews, inbound referrals, and outbound links. To protect user privacy, IRDI will activate all available privacy protection features of the web analytics tool (for example, IP anonymization).

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

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

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

This project will create an application prototype for a research data index, built using a LAMP stack (Linux, Apache, MySQL, and PHP). All metadata will be created and distributed using open text file types including CSV and JSON-LD. We will also create a package installer (DMG, EXE, or PKG). Documentation and training materials will be available on our website as HTML and accessible PDF documents. Conference presentation will be made available via Zenodo.

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 application prototype for a research data index will be built using a LAMP stack (Linux, Apache, MySQL, and PHP), with PHP and Javascript (including the jQuery library) scripts. Equipment required will be personal computers and servers at the MSU Library.

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).

Application LAMP stack (Linux, Apache, MySQL, and PHP). Package installer (DMG, EXE, or PKG). All metadata will be created and distributed using open text file types including CSV and JSON-LD. Documentation and training materials (HTML and accessible PDF).

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

We will conduct usability checks throughout the process. In-progress project will be presented at Code4Lib for feedback from the Library developer community. User feedback will be solicited via user forums and user feedback forms.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

The MSU Library has robust digital preservation infrastructure, including monthly local backups, nightly differentials, and 3x-yearly geographically-distributed backups in Ohio and Texas. Supported formats are assessed for migration every 5 years. For full digital preservation policies and procedures, please see http://www.lib.montana.edu/archives/digital-preservation/

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).

The project team has built a preliminary metadata model for the IRDI prototype informed by Schema.org, Data Catalog Vocabulary (DCAT), DataCite, DATS, and Project Open Data metadata schemas. The final prototype will be made available in Github, including a readme and instructions for use. All metadata will be created and distributed using open text file types including CSV and JSON-LD. We will use the Zenodo/Github integration to provide a DOI and Dublin Core descriptive metadata for the prototype code.

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

The Zenodo/Github integration will preserve our metadata for the long term. Zenodo is fully run on open source products, with technical, security, and preservation features that follow community best practices. Data files and metadata are backed up nightly and replicated into multiple copies. All data files are stored with a MD5 checksum, and files are regularly checked against their checksums to ensure that file content remains constant. Data retention is guaranteed for the lifetime of the repository. In case of closure of the repository, best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories (see http://about.zenodo.org/infrastructure/ and http://about.zenodo.org/policies/).

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).

Metadata from the IRDI prototype will be created and distributed using open text file types including CSV and JSON-LD, and will be available via RSS and API. IRDI metadata will be optimized for discovery by commercial search engines.

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).

Our Github repository and supporting materials will be made available in Zenodo (https://zenodo.org), a data repository hosted by CERN—a memory institution for High Energy Physics, known for its pioneering work in Open Access. Zenodo is funded by CERN, the European Commission via OpenAire projects, and the Alfred P. Sloan Foundation. All open content in Zenodo is available for the public to view and download, free of charge, via an online interface. Data depositors may opt to embargo their content for a limited period of time before it is made public, or they may opt to restrict their data, approving access only to requesters who meet certain requirements. Zenodo ascribes to the FAIR Guiding Principles for scientific data management and stewardship, aiming to ensure that data is findable, accessible, interoperable, and reusable (http://about.zenodo.org/principles/).

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.

Selected examples:
MSU Library Github repository https://github.com/msulibrary/
MSU Research Citations App https://github.com/msulibrary/msu-research-citations
Presentation slides by Sara Mannheimer et al. https://zenodo.org/record/1038123
Dataset by Sara Mannheimer et al. https://doi.org/10.15788/m2059z

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.

The software will be an application prototype for a research data index, built using a LAMP stack (Linux, Apache, MySQL, and PHP). We will also create a package installer (DMG, EXE, or PKG). The prototype will harvest metadata from third-party data repositories to provide an index of institutional research datasets. The primary audience are small and mid-sized institutions who would like to promote discovery and access to research data while expending fewer resources than would be required to build and maintain an institutional data repository.

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.

IRDI will build on related projects such as DataMed, NYU Health Sciences Library Data Catalog, Cinergi, OmicsDI, and SHARE, offering three critical innovations: (1) IRDI brings an institutional focus to the automated collection of metadata from external data repositories. Automated metadata collection allows the index to be populated with metadata for institutional datasets with less manual effort from library employees and therefore less resource expenditure from the institution; (2) IRDI will also automatically generate new descriptive metadata for individual datasets using external topic mining of scholarly profile sources like ORCID and Google Scholar Profiles; and (3) IRDI will be optimized for discovery through leading commercial search engines.

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.

We will build the prototype using a LAMP stack (Linux, Apache, MySQL, and PHP). We chose these languages and platforms because they are the basis of our digital library software. The IRDI prototype will be a reworked version of that existing software. The IRDI prototype is based on a system developed at MSU Library that uses Javascript and PHP scripts to parse RSS feeds and harvest metadata for institutional publications. Building off of existing systems at MSU will allow us to be more efficient in terms of training and systems.

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

As a metadata index, IRDI's key interoperation will be with its metadata. The project team has experimented with some early prototyping of the IRDI system, including building a preliminary metadata model informed by Schema.org, Data Catalog Vocabulary (DCAT), DataCite, DATS, and Project Open Data metadata schemas. We will also be in contact with an advisory board that will include personnel from DataMed, NYU Health Sciences Library Data Catalog, and other data discovery projects. This collaboration will help ensure interoperability between data discovery projects.

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

The IRDI prototype requires a LAMP stack (Linux, Apache, MySQL, and PHP).

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

We will create a project website with links and information about IRDI, installation documentation, and FAQs. IRDI code will be given a Creative Commons Attribution (CC BY) license and released via Github, and IRDI metadata will be available openly via RSS feed and an API. A package installer will be provided so that small and mid-sized institutions with limited developer expertise can easily install the system. The project team will also host training webinars to guide users through installation and implementation of the IRDI prototype.

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

Selected examples:
Catnip app: https://www.lib.montana.edu/catnip/
Research publications metadata harvesting app: http://www.montana.edu/research/publications/ (for more information, please see https://doi.org/10.5860/crl.78.7.952).

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.

The IRDI prototype code will be assigned a GNU General Public License. All other digital products will be assigned Creative Commons Attribution (CC BY) licenses. GNU GPL and CC BY are minimally restrictive licenses, allowing for maximum reuse, remixing, and redistribution.

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

The final prototype will be made available in Github, including a readme and instructions for use. We will use the Zenodo/Github integration to provide a DOI and Dublin Core descriptive metadata for the prototype code. To promote and disseminate the prototype, we will create a project website with links and information about IRDI, installation documentation, and FAQs. A package installer will be provided so that small and mid-sized institutions with limited developer expertise can easily install the system. The project team will also host training webinars to guide users through installation and implementation of the IRDI prototype.

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

Name of publicly accessible source code repository: Github and Zenodo

URL https://github.com/msulibrary, https://zenodo.org

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.

N/A

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?

N/A

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).

N/A

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.

N/A

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).

N/A

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?

N/A

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

N/A

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

Name of repository: N/A

URL: N/A

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

This plan will be reviewed at the beginning of each the project (Phase 1—Foundational activities, October 2018; Phase 2—Prototype startup, December 2018; Phase 3—Troubleshooting and cleanup, June 2019; Phase 4.—Dissemination of results, August 2019).