A Prototype for the Institutional Research Data Index

Abstract: The Montana State University (MSU) Library seeks a $25,000 Sparks Grant in the Curating Collections project category to build a prototype for the Institutional Research Data Index (IRDI), a metadata index that promotes discovery and reuse of institutional research datasets by automatically harvesting metadata from third-party data repositories and by optimizing its contents for commercial search engines.

Statement of National Need: Academic libraries are increasingly participating in research data publishing 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). In addition, many institutional repository 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 (for instance, Dryad Digital Repository still uses DSpace 1, a version first released in 2002). 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—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. In order to support research data programs and repositories, libraries are either increasing spending by buying vendor solutions, or replicating work by building and managing individual instances of data repository software.

Project Design: We propose a Sparks Grant to build a prototype for the 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. IRDI will build on similar projects such as NYU Health Sciences Library Data Catalog, NIH DataMed, and SHARE, offering two critical innovations: (1) the automated collection of metadata, which 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; and (2) search engine optimization to ensure that data index content is easily discoverable 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, DataCite, DATS, and Project Open Data metadata schemas. 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 to harvest metadata for institutional publications (Sterman & Clark, 2017). IRDI code will be open source, and IRDI content will be available via RSS feed and an API. The insights gained from building IRDI will be disseminated across our community via conference presentations, peer-reviewed journal articles, and a webinar.

Potential Challenges: 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. We have investigated querying data repositories for the full names of MSU researchers, but this strategy is labor-intensive and fails to account for name disambiguation. ORCID usage is increasing, and could provide a partial solution; the data repository Zenodo is one notable adopter of ORCID integration. A second

complexity is that 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. Fortunately, there are third-party data repositories 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 the cost of archiving unique research data in such repositories.

National Impact: The IRDI pilot will have two 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)). Second, IRDI will reinforce the idea that research data is a legitimate scholarly product: it will interoperate with institutional Research Information Management (RIM) systems, it will function as a database that institutions can query for assessment purposes, and it will create a centralized, public interface where institutions can showcase all of the research data published by institutional researchers. Additionally, a third, much broader impact can be realized if the IRDI pilot is successful: IRDI can be expanded into a single, unified index for academic institutional research data 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 data. 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.

Relevance to Project Category: The IRDI prototype aims to provide increased access to research data, and to encourage use and reuse of research data. If successful, the prototype can scale up into a regional or national shared service that provides access to research data for the research, educational, and public communities. This expanded system would create alliances and networks of libraries surrounding discovery of institutional research data. IRDI will be especially useful for small and mid-sized academic libraries with limited resources.

Project Staff and Partners: Our team consists of three MSU Library employees. Project Director Sara Mannheimer, Data Librarian, has expertise in data repository development and data discovery. Key project staff Jason Clark, Head of Special Collections & Archival Informatics, has expertise in metadata and semantic web. Key project staff Jim Espeland, Software Engineer, has expertise developing library applications and systems. The grant will fund an additional student developer to build the IRDI system, under Espeland's mentorship. We have also established a partnership with the MSU Office of Planning and Analysis—which manages the MSU RIM system— to ensure that IRDI is interoperable with RIM systems. Throughout the project, we will keep in communication with the data curation community, including related data discovery projects such as the aforementioned NYU Data Catalog, DataMed, and SHARE; 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).

Budget: The $25,000 grant will be allocated as follows: $6,525 in salary and benefits support for Mannheimer (2.5%), Clark (1%), and Espeland (5%); $12,000 to fund a paid internship for a computer science student, who will act as the IRDI software developer under Espeland's mentorship; $3,192 in conference travel to promote the project at Code4Lib 2019; and $3,285 in indirect costs.