Journal of eScience Librarianship

Volume 10
Issue 1 *Special Issue: Research Data and Preservation (RDAP) Summit 2020*

Article 3

2021-01-19

Dataset Search: A lightweight, community-built tool to support research data discovery

Sara Mannheimer
Montana State University

Et al.

ISSN 2161-3974

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Dataset Search: A lightweight, community-built tool to support research data discovery

Sara Mannheimer, Jason A. Clark, Kyle Hagerman, Jakob Schultz, and James Espeland

Montana State University, Bozeman, MT, USA

Abstract

Objective: Promoting discovery of research data helps archived data realize its potential to advance knowledge. Montana State University (MSU) Dataset Search aims to support discovery and reporting for research datasets created by researchers at institutions.

Methods and Results: The Dataset Search application consists of five core features: a streamlined browse and search interface, a data model based on dataset discovery, a harvesting process for finding and vetting datasets stored in external repositories, an administrative interface for managing the creation, ingest, and maintenance of dataset records, and a dataset visualization interface to demonstrate how data is produced and used by MSU researchers.

Conclusion: The Dataset Search application is designed to be easily customized and implemented by other institutions. Indexes like Dataset Search can improve search and discovery for content archived in data repositories, therefore amplifying the impact and benefits of archived data.

Correspondence: Sara Mannheimer: sara.mannheimer@montana.edu

Received: June 4, 2020 **Accepted**: September 8, 2020 **Published**: January 19, 2021

Data Availability: Code associated with this paper is available in Zenodo, via Github at: https://doi.org/10.5281/zenodo.4046567. MSU Dataset Search is available at: https://arc.lib.montana.edu/msu-dataset-search.

Disclosures: The authors report no conflict of interest. The substance of this article is based upon a lightning talk at RDAP Summit 2020. Additional information at end of article.

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

Introduction and Background

Sharing the scientific data that underlie results is increasingly seen as a vital part of scholarly communication (Baker 2017; Boulton et al. 2012). Sharing research data has multiple potential benefits. Shared data can increase time efficiency and cost efficiency by allowing researchers to reuse data rather than collect new data (Pronk 2019); it can support reproducibility and replicability for scientific research (National Academies of Sciences, Engineering, and Medicine 2019); it can produce new discoveries to advance science (Fienberg et al. 1985); it can increase visibility and impact of research (Piwowar and Vision 2013); encourage new, mutually-beneficial collaborations between researchers (Pasquetto, Borgman, and Wofford 2019); and shared data can be used in the classroom and during apprenticeships to support the next generation of researchers (Haaker and Morgan-Brett 2017; Kriesberg et al. 2013).

In the United States, research data that result from public funding are further considered to be a public asset that should be shared openly (Holdren 2013). In response to this idea, federal funding agencies now require sharing data with other researchers. The National Science Foundation's policy states, "grantees are expected to encourage and facilitate [data] sharing" (National Science Foundation 2011); and the National Institutes of Health suggest that "data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data" (National Institutes of Health 2003). An increasing number of scientific journals also require that researchers share the data underlying their published articles. In 2011, a group of journals in the field of evolution coordinated to implement the Joint Data Archiving Policy requiring authors to publish the data underlying their publications (Dryad Digital Repository 2011), and other scientific journals have followed suit, including PLOS journals (PLOS 2014) and the Committee of Medical Journal Editors (Taichman et al. 2017).

Researchers share their data in multiple ways: as supplementary material to published articles, as downloads on institutional or personal websites, through archiving in data repositories, or by sharing data "upon request"—that is, in response to inquiries from other researchers (Kim and Stanton 2016; Tenopir et al. 2015; Wallis, Rolando, and Borgman 2013). The 2016 FAIR Data Principles propose that beyond just being shared, data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). From a data stewardship perspective, and in order to best support FAIR data, sharing via data repositories allows for the most reliable long-term discovery, access, and preservation for shared data (Witt 2008; Poole 2015; Kim and Zhang 2015). Data repositories also integrate into the scholarly communication ecosystem, supporting data citation practices for data creators (Nature Biotechnology 2009; Fenner et al. 2019). Therefore, data repositories are often the preferred method for data sharing—for example, as stated the PLOS data sharing policy (Federer et al. 2018). As of May 2020, the Registry of Research Data Repositories (re3data.org) has indexed 1068 unique repositories in the United States (Registry of Research Data Repositories

2020). These data repositories can be categorized into four key types (Pampel et al. 2013):

1. Institutional research data repositories, which are often operated by academic libraries—e.g. the University of Michigan's Deep Blue Data (University of Michigan 2020) and Harvard's Dataverse (Harvard College 2020);

2. Disciplinary research data repositories that archive data in specific formats or from specific subjects—e.g. GenBank (Benson et al. 2013) and the Qualitative Data Repository (Center for Qualitative and Multi-Method Inquiry 2020);

3. Multidisciplinary research data repositories with broader collecting missions—e.g. figshare (Digital Science & Research Ltd 2020), Zenodo (CERN Data Centre 2020), and Dryad Digital Repository (Dryad 2020);

4. Project specific research data repositories—e.g. the National Snow and Ice Data Center (NSIDC 2020).

Data repositories are still a relatively new development in scholarly communication, and their infrastructure and metadata are far less standardized than in scientific journals—for instance, data repositories don't always require that depositors add institutional affiliation, and metadata are also often entered by the depositor, rather than entered in a standardized way by professional catalogers (Marcial and Hemminger 2010). In 2010, Marcial and Hemminger also identified preservation as an issue; only 62% of the data repositories they surveyed had "a clear mention of a preservation policy or similar" (2038). However, an increasing number of data repositories are now certified under initiatives such as the CoreTrustSeal Trustworthy Data Repositories Requirements, a set of standards for data stewardship that certify that repositories support healthy infrastructure and long-term preservation for repositories (CoreTrustSeal 2020). Additionally, the TRUST Principles can help repositories become more trustworthy data stewards, and help researchers select a trustworthy repository for data sharing (Lin et al. 2020).

While data repositories are increasingly focusing on long-term data stewardship, they still have room to grow in terms of promoting discovery for their resources. A 2017 study of natural resources and environmental scientists found that "while institutional repositories were commended by interviewees for providing permanent archiving and long-term preservation, for supporting storage and download, and for ensuring accessibility and credibility… [they were] not particularly valued for searchability and discoverability" (Shen 2017, 120). While efforts have been made to improve discovery for institutional repositories (Arlitsch and O'Brien 2012), Mannheimer, Sterman, and Borda (2016) find that research data are discovered and reused most often if they are: (1) archived in disciplinary research data repositories; and (2) indexed in multiple online locations.

An increasing number of recent projects focus on indexing data in repositories, including the NIH-funded DataMed (Chen et al. 2018), which uses the DATS suite of tags to support automatic indexing of scientific datasets (Sansone et al. 2017); SHARE, which cooperates with institutional repositories to use "a schema-agnostic approach" to metadata aggregation (Hudson-Vitale et al. 2017, sec. 1, para. 6); Elsevier DataSearch, which uses a two-tiered word embedding analysis to match natural language queries and a formal ontology assignment (Scerri et al. 2017); and Google Dataset Search, which uses Schema.org as a unifying metadata schema and which came out of beta in 2020 (Noy 2020). However, dataset indexing projects such as these may not reveal all available research data. Some research data cannot be published openly in data repositories, either because the research is still in-progress, or because the data are sensitive in nature. This has motivated the creation of data catalogs that include restricted data. Notable projects are NYU Langone Health Sciences Library's Data Catalog (Lamb and Larson 2016) and its fellow members of the Data Discovery Collaboration (formerly the Data Catalog Collaboration Project) (Read et al. 2018).

The Montana State University (MSU) Library aims to bring together ideas from each of the projects described above, as well as some innovations, to encourage discovery and reuse of datasets from MSU researchers.

The Montana State University Dataset Search

Montana State University (MSU) is a mid-sized university. In the 2019-2020 academic year, the university had 16,766 students (Montana State University 2020) and 56 library employees (MSU Library 2020). In 2019, MSU Library joined Dryad Digital Repository as an institutional member to support trustworthy, long-term preservation for research datasets at our institution. This allows us to focus our local efforts on research data curation and discovery. As part of these efforts, we built a Dataset Search tool to support discovery, access, and reuse for research datasets from our institution (Mannheimer et al. 2019; MSU Dataset Search 2020).

MSU Dataset Search[1] is a lightweight, open source, scalable, sustainable, and standardized search tool that indexes datasets created by MSU researchers that are archived in public data repositories. The project has been funded by the Institute of Museum and Library Services (Montana State University 2018), and by the National Network of Libraries of Medicine, Pacific Northwest Region (Mannheimer and Clark 2019). Unlike a data repository, MSU Dataset Search does not archive research datasets themselves. Instead, it harvests metadata from third-party data repositories that archive research datasets, and serves the metadata via an online interface[2]. MSU Dataset Search project joined several other institutions in the Data Discovery Collaboration (DDC 2020) in 2020. We have also partnered with the MSU Center for American Indian and Rural Health Equity

1 https://arc.lib.montana.edu/msu-dataset-search

2 Dataset Search code is based on a similar project at Montana State University that harvests metadata records for text-based publications (see Sterman and Clark 2017).

(CAIRHE 2020) to support a pilot effort to manually produce metadata records for restricted datasets that can be accessed by contacting the Center. Indexing these datasets supports research transparency and data discovery and access for the Center's community stakeholders.

MSU Dataset Search complements existing data discovery efforts by indexing and creating metadata records for data in repositories, showcasing the data created at our institution through a visualization dashboard, as well as by creating metadata records for restricted data. MSU Dataset Search also adds three innovative features to these efforts. First, Dataset Search brings an institutional focus to the automated collection of metadata from third-party data repositories; automated metadata collection allows the index to be populated with metadata for local research datasets with less manual effort from library employees and therefore less resource expenditure from the institution. Second, Dataset Search is optimized for commercial web search engines, which supports discovery of MSU datasets on the open web. Third, Dataset Search automatically generates new descriptive metadata for individual datasets using external topic mining of scholarly profile sources like ORCID and Google Scholar Profiles.

Building the Tool

To begin building the Dataset Search tool, the team needed to understand how to identify datasets that had been published by researchers at our institution. Centering this question led us to also think about how we could construct the tool to allow other institutions to apply the software. In moving from our specific use case to a broadly-applicable model, five components became core features of the application: a streamlined browse and search interface, a data model based on dataset discovery, a harvesting process for finding and vetting datasets stored in external repositories, an administrative interface for managing the creation, ingest and maintenance of dataset records, and a dataset visualization interface to demonstrate how data is produced and used by our researchers. These components are discussed in more detail below.

Browse and search interface

The need for an interface to allow for search and retrieval was a primary consideration. The team wanted a clean interface that made it easy for users to search, browse, and access datasets in external repositories. In the section "Lessons learned and continued challenges," we further discuss the particular challenge of designing the interface and our work with a designer to come up with primary actions for the application. These discussions helped us isolate the fundamental user experience; our team focused on helping users identify the purpose of the application, find a particular dataset, and then link from the metadata in our system to the repository where the dataset is stored. These core actions define the primary interface. The visual layout for the Dataset Search landing page can be seen in figure 1.

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

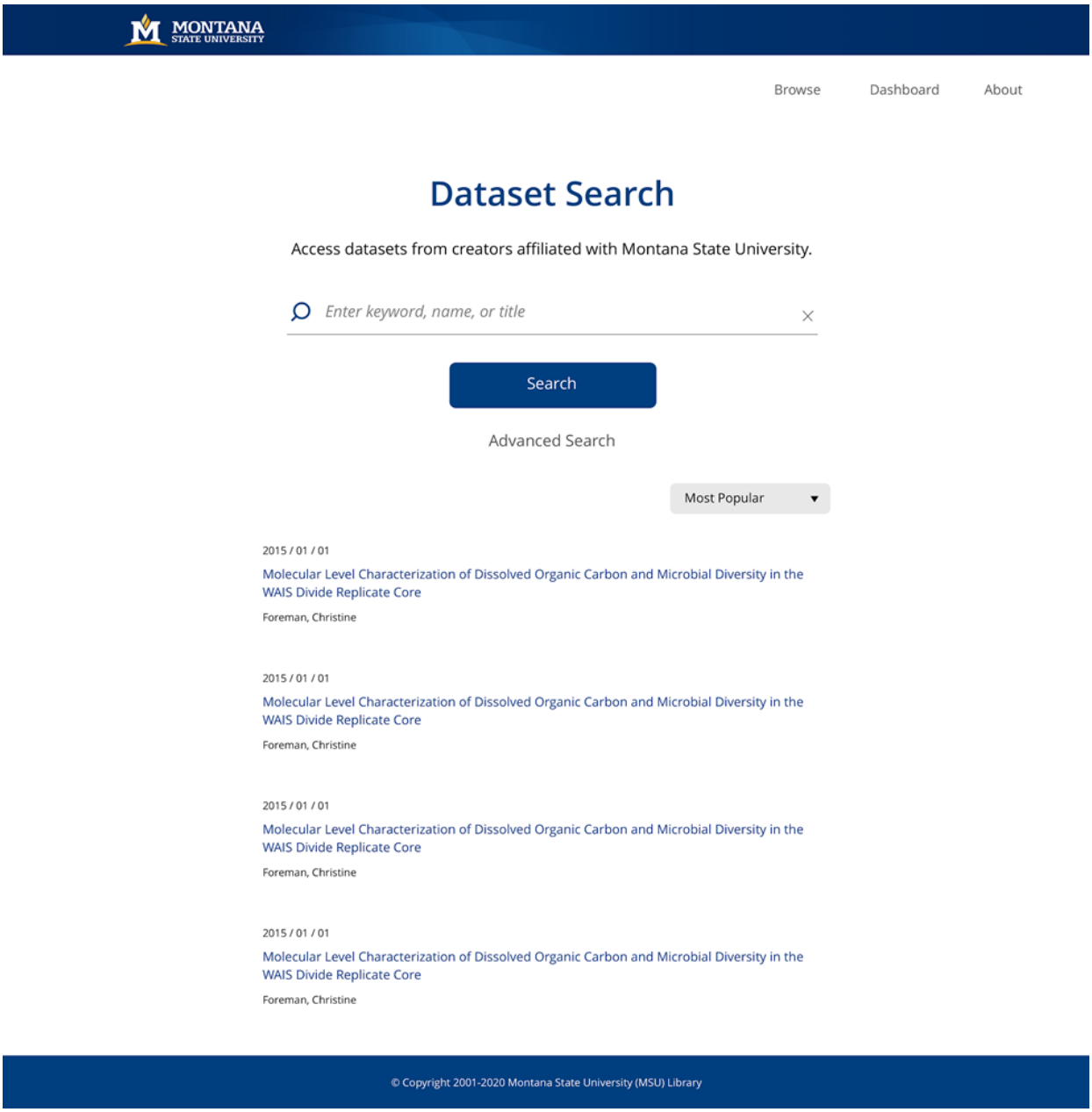

Figure 1: MSU Dataset Search, Home Page

A user is able to recognize quickly the reason for their being on the page is to search for datasets. In turn, the search box and list of recent datasets are calls to action that impart what next steps might be, but also indicate that a user is at the landing page of the Dataset Search application. The landing page clearly directs the user to search and browse through the system.

Beyond the landing page and search/browse results, a user is led into a view of item metadata that displays a title and description, a permanent identifier for dataset, and a button linking to the actual dataset in an external data repository.

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

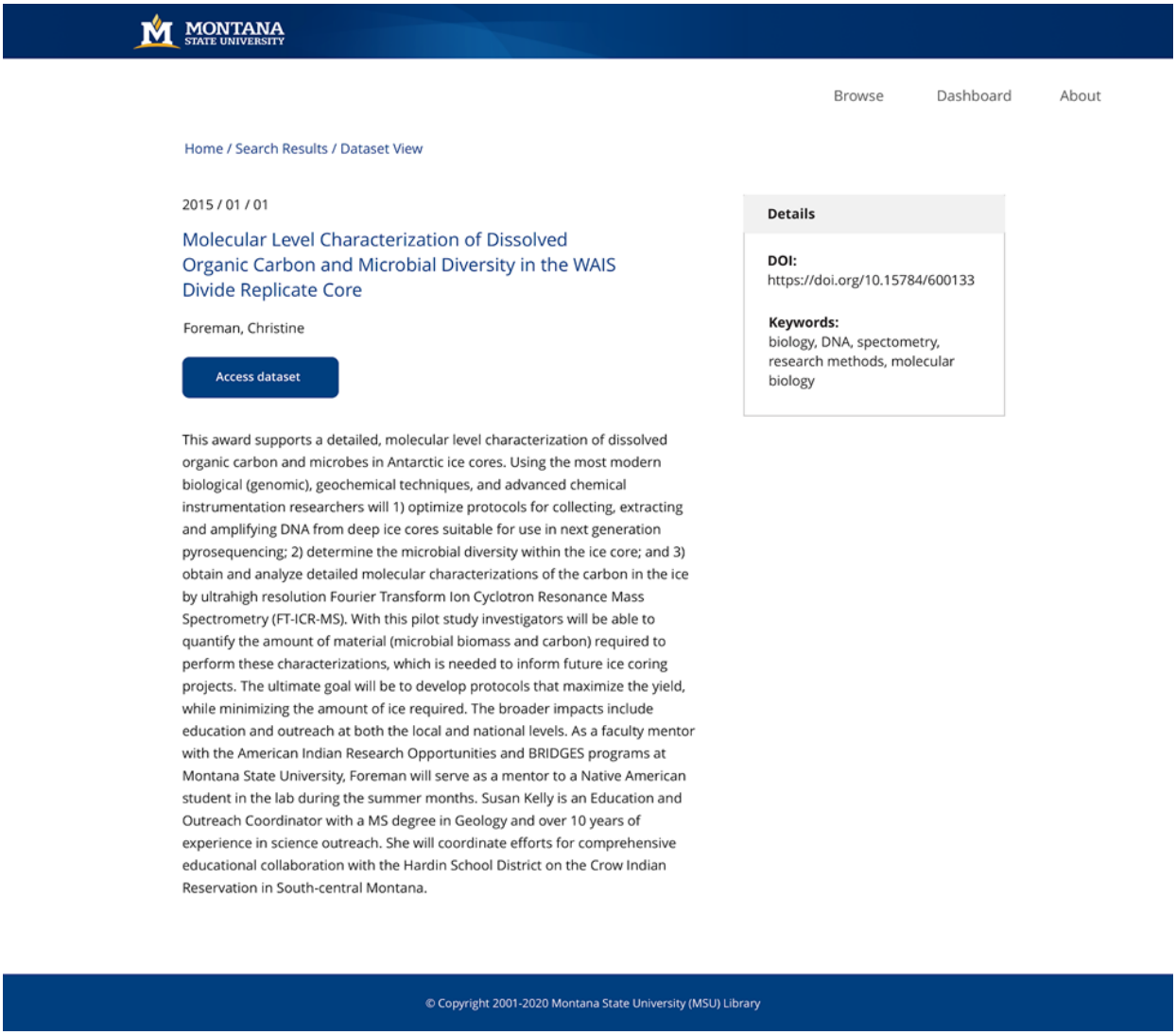

Figure 2: MSU Dataset Search, Item Page

The item page is the link between the local metadata record and the external repository that provides access to the dataset. The metadata on the item page also allows us to catalog MSU researchers and the types of data they produce.

Data model based on dataset discovery

The data model for the datasets was also essential. Our research revealed no shortage of metadata schemes to follow. We ultimately took our cues from the Google Dataset Search metadata, which applies the Schema.org web vocabulary. This structured data vocabulary is a widely-adopted standard, and it sets up a series of types and properties to describe the datasets with a goal of indexing for discovery in commercial search engines (Schema.org 2020a). The overarching goal of discovery suited our needs, but there were times where the data model

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

needed some enhancement for administrative and technical metadata. Schema.org prioritizes the "aboutness" of the dataset which leads to primary properties that help a person understand more about the content within the dataset. Properties like measurementTechnique (Schema.org 2020b) and variableMeasured (Schema.org 2020c) are just two examples of this "aboutness" prioritization within Schema.org. Within our data model, we made additions to support linked data identifiers and we added administrative properties like `dataset_urlHash`, `recordInfo_recordContentSource`, `dataset_conditionsOfAccess`. An example of our primary entity table, a datasets table, is featured in the figure below.

```
CREATE TABLE IF NOT EXISTS `datasets` (
  `recordInfo_recordIdentifier` int(10) NOT NULL COMMENT 'record id',
  `dataset_name` varchar(300) NOT NULL DEFAULT '' COMMENT 'dataset title',
  `dataset_doi` varchar(300) DEFAULT NULL COMMENT 'original dataset DOI, points at external record',
  `dataset_repositoryName` varchar(255) DEFAULT NULL COMMENT 'name of repository',
  `dataset_funder` varchar(255) DEFAULT NULL COMMENT 'INBRE or NIH',
  `dataset_url` varchar(300) DEFAULT NULL COMMENT 'direct url for the actual dataset content',
  `dataset_description` text COMMENT 'dataset abstract',
  `dataset_keywords` varchar(255) DEFAULT NULL COMMENT 'dataset comma-delimited content keywords',
  `dataset_temporalCoverage` varchar(30) DEFAULT NULL COMMENT 'date dataset published e.g., 1950-01-01/2013-12-18',
  `dataset_spatialCoverage` varchar(30) DEFAULT NULL COMMENT 'geoshape box coordinates OR latitude/longitude',
  `dataset_category1` varchar(255) DEFAULT NULL COMMENT 'linked data category',
  `dataset_category1_uri` varchar(255) DEFAULT NULL COMMENT 'linked data URI',
  `dataset_category2` varchar(255) DEFAULT NULL COMMENT 'linked data category',
  `dataset_category2_uri` varchar(255) DEFAULT NULL COMMENT 'linked data URI',
  `dataset_category3` varchar(255) DEFAULT NULL COMMENT 'linked data category',
  `dataset_category3_uri` varchar(255) DEFAULT NULL COMMENT 'linked data URI',
  `dataset_category4` varchar(255) DEFAULT NULL COMMENT 'linked data category',
  `dataset_category4_uri` varchar(255) DEFAULT NULL COMMENT 'linked data URI',
  `dataset_category5` varchar(255) DEFAULT NULL COMMENT 'linked data category',
  `dataset_category5_uri` varchar(255) DEFAULT NULL COMMENT 'linked data URI',
  `dataset_encodingFormat` varchar(30) DEFAULT NULL COMMENT 'dataset format type e.g., CSV',
  `dataset_license` varchar(255) NOT NULL DEFAULT 'Attribution Non-Commercial Share Alike Creative Commons ' COMMENT 'dataset
copyright conditions',
  `dataset_conditionsOfAccess` varchar(255) DEFAULT NULL,
  `dataset_version` varchar(30) DEFAULT NULL COMMENT 'dataset version number',
  `dataset_sameAs` varchar(300) DEFAULT NULL COMMENT 'dataset duplicate content URL for disambiguation if/when dataset is listed in
multiple repositories',
  `dataset_urlHash` varchar(40) DEFAULT NULL COMMENT 'sha1 hash of harvest info to help with deduping during harvest',
  `dataset_relatedMaterial` varchar(255) DEFAULT NULL COMMENT 'list any related material include DOIs',
  `recordInfo_languageOfCataloging` varchar(5) NOT NULL DEFAULT 'en' COMMENT 'language of record',
  `recordInfo_recordContentSource` varchar(10) NOT NULL DEFAULT 'MZF' COMMENT 'oclc institution id',
  `recordInfo_recordCreationDate` date NOT NULL DEFAULT '0000-00-00' COMMENT 'date record created',
  `recordInfo_recordModified` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP COMMENT 'date record
modified',
  `status` varchar(10) CHARACTER SET ucs2 NOT NULL DEFAULT 'u' COMMENT 'record activity status'
) ENGINE=MyISAM DEFAULT CHARSET=utf8;
```

Figure 3: Data Fields for Describing 'Datasets'

Figure 3 gives a picture of the 'datasets' table as a SQL CREATE query, but it also demonstrates where parts of the discovery metadata are not enough. Access restrictions (`dataset_conditionsOfAccess`), dataset sources (`recordInfo_recordContentSource`), and methods for deduplicating datasets (`dataset_urlHash`) were added to supplement and build metadata to support technical and administrative tasks.

To enrich our data model, we have chosen to provide as much information about authors as possible. Currently, this means we are scraping Google Scholar (Google 2020a) for MSU's faculty profiles via Python script for their posted keywords. This script then takes the keywords, cross references them with WikiData (Wikidata 2020) and grabs the relevant machine label. Keyword and machine label are then stored side by side in the database. This means authors can be linked via their interests or professional skills and their published works can be found in a single query of our database.

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

Harvesting process for finding and vetting datasets

In building the tool, the team also set requirements around harvesting and vetting datasets for inclusion in the MSU inventory. This was in many ways the central organizing principle for the application. We needed to create a software process to search multiple, external dataset repositories and identify datasets that are affiliated with MSU research or produced by MSU researchers. A number of options from web scraping of search result pages to Application Programming Interface (API) querying were considered. Our team settled on API querying as it allowed an explicit contract between our application and the external dataset repositories as well as a structured data response that we could write a software process to consume.

```
<feed>
	<publisher>Dryad</publisher>
	<entryTag>data</entryTag>
	<feedStructure>
		<root>
			<data>
				<id>uid</id>
				<attributes>
					<identifiers>
						<identifier>link</identifier>
					</identifiers>
					<creators>
						<name>creator</name>
					</creators>
					<titles>
						<title>title</title>
					</titles>
					<descriptions>
						<description>description</description>
					</descriptions>
				</attributes>
			</data>
		</root>
	</feedStructure>
</feed>
<feed>
	<publisher>DataCite-Dryad</publisher>
	<entryTag>data</entryTag>
	<feedStructure>
		<root>
			<data>
				<id>uid</id>
				<attributes>
					<identifiers>
						<identifier>link</identifier>
					</identifiers>
					<creators>
						<name>creator</name>
					</creators>
					<titles>
						<title>title</title>
					</titles>
					<descriptions>
						<description>description</description>
					</descriptions>
				</attributes>
			</data>
		</root>
	</feedStructure>
</feed>
```

Figure 4: Example XML mapping for an individual API

Currently, the MSU Dataset Search tool has functionality for storing XML feeds or API responses that are available for consumption from data repositories. When a feed is selected and added to the application, a PHP script breaks down the feed and determines the repeating tag used to store entries. There are no formal

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

guidelines for how these repositories structure their feeds, so there are not any normalized naming schemes we can rely on. However, the repeating tag will always be the tag in a feed with the highest product between the number of instances it appears and the number of children it has. With the help of a curator using an HTML form, we can identify the tags in the feed as we have named them in our database and form an XML map of the feed.

Using the extracted XML map, we can traverse any feed according to its structure and auto-populate records to be inserted into the database. Should a feed ever change, we can either update the file containing the XML map, or re-add the feed and the script will find the corresponding tags again. By automating this process, we can handle a variety of different feed structures and tag naming conventions.

Beyond the initial querying and harvest of our datasets through the APIs, we needed a way to vet and deduplicate our dataset records. The team settled on a deduplication string that is currently a combination of the dataset title, link, description, creator, pubDate, and uid (if they are set). This is then used to create the dataset_urlHash which is a unique identifier that we can check against to verify if we have already harvested a dataset record. The team is encouraged by the results here as it allows us to automatically check for duplicates and has increased the efficiency of our ingest process.

Administrative interface

With the data model and harvesting in place, we needed a secondary interface that would allow us to manage the data. We constructed a series of web forms to enable harvesting, adding, updating, and deleting of metadata.

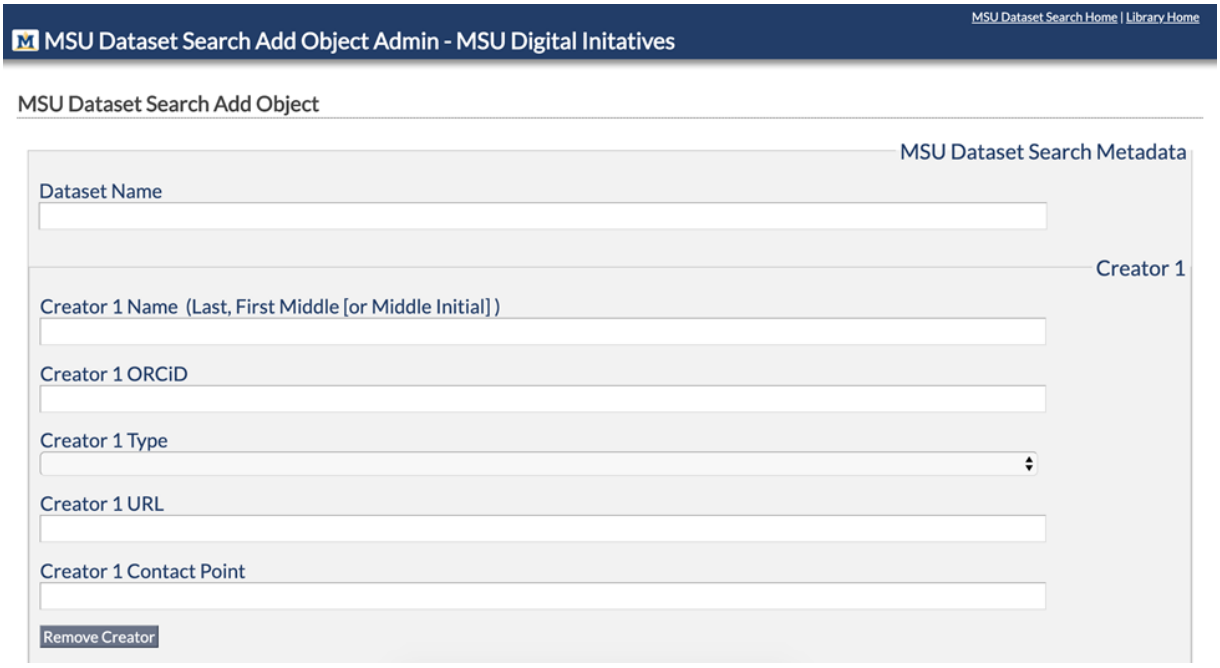

Figure 5: MSU Dataset Search, Metadata Entry Form

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

The administrative interface also includes our harvesting routine for automatically populating our dataset records from external sources. This view is an editing table that pulls in data from theses external sources and then allows a curator to review or accept a dataset as a record for MSU Dataset Search. The view below shows the table as it is being populated.

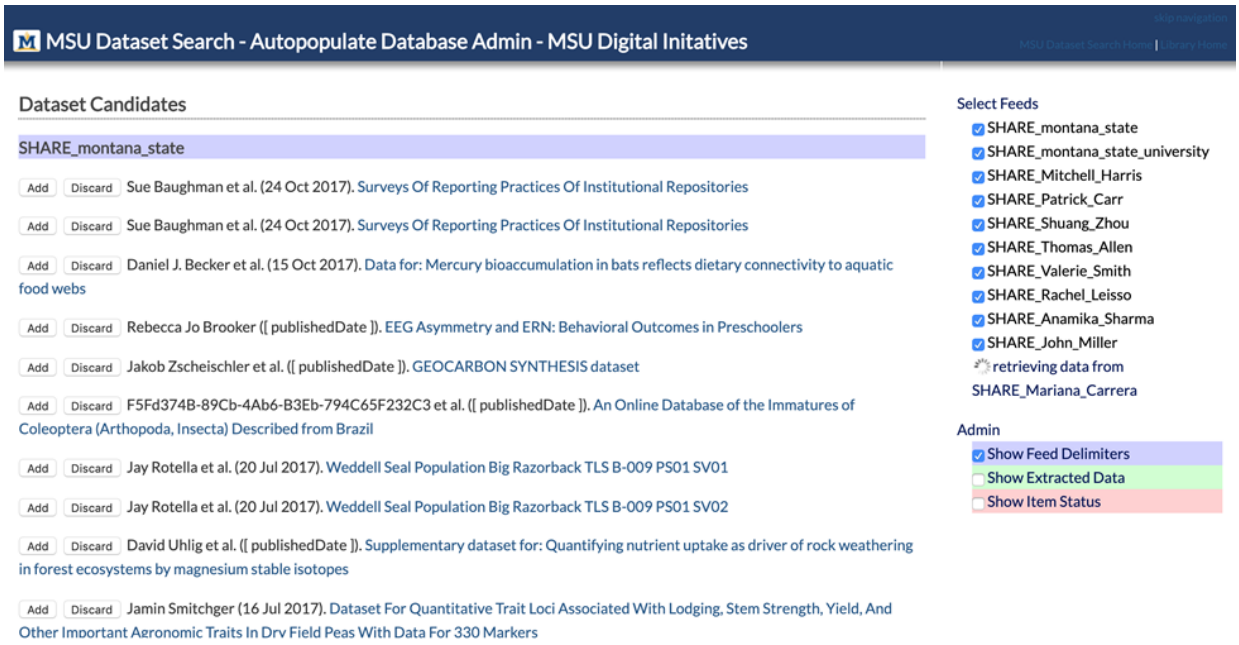

Figure 6: MSU Dataset Search, "Autopopulation" Harvest View

The harvesting view also allows the curator to control the amount of metadata that is visible and helps create a minimum viable metadata record that provides our catalog with an automated routine for data entry.

The administrative interface can also be used to manually create metadata records. Partnering with the Center for American Indian and Rural Health Equity (CAIRHE), our team has created pilot records in the system to promote discovery for restricted datasets. Instead of linking to the dataset in an external repository, the system directs users to contact the Center to request the dataset.

Visualization interface to demonstrate how data is produced and used

Part of our goal in creating the Dataset Search was to showcase research and research data at our institution. Our team considered how public dashboards could help shape different views and understandings of our dataset inventory, providing quick snapshots, trends, and analysis of the datasets in the application. These data dashboards are currently in-development.

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

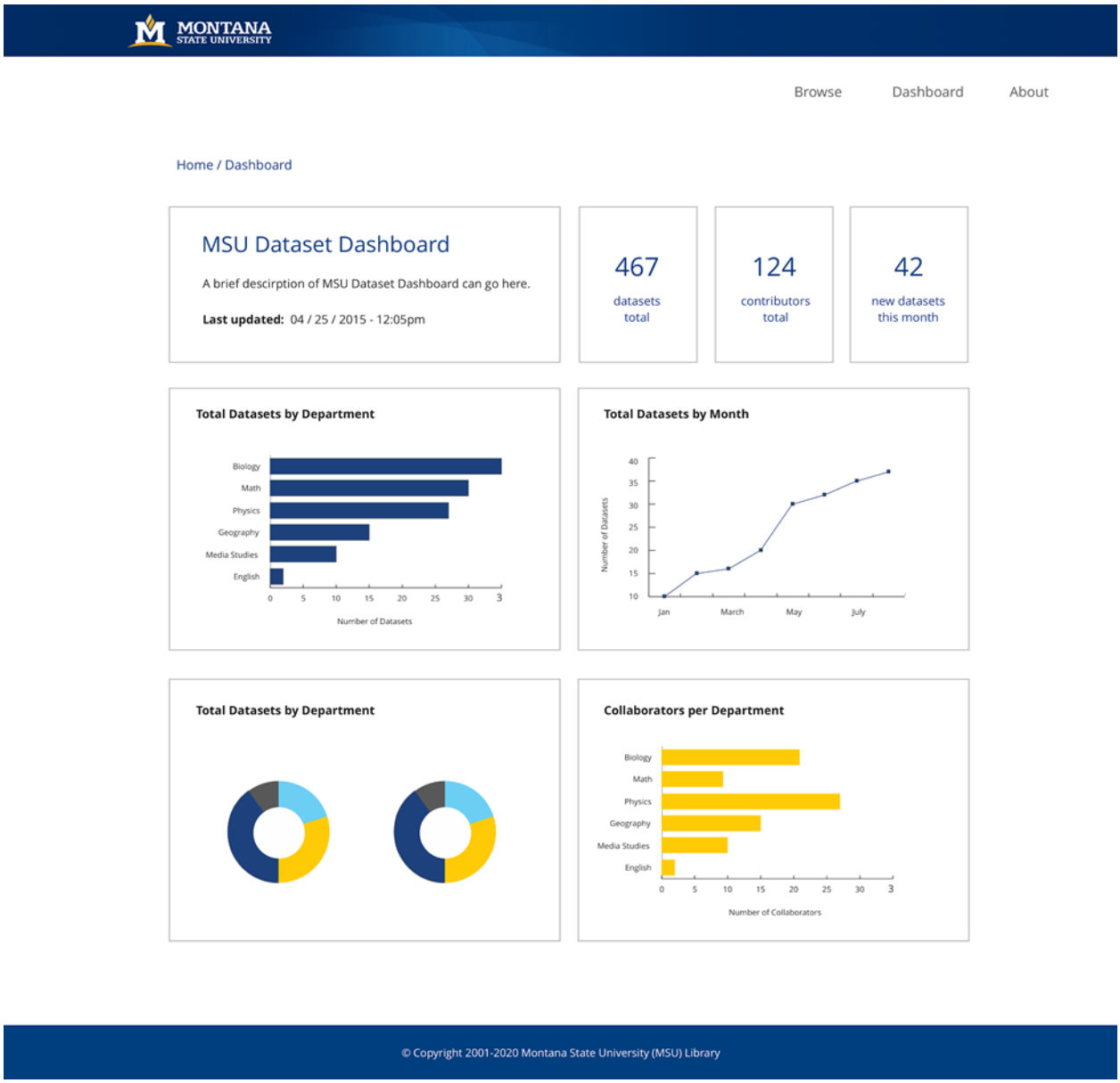

Figure 7: MSU Dataset Search, Public Dashboard View

There are a variety of fields that we capture within our database that allow a user to filter metadata by certain fields. To visually capture this, we have a series of queries that will display current data as infographics using D3.js, a Javascript visualization library. We are working to prototype dashboard landing pages unique to each field a user may want to filter on such as: author, college, department, affiliation, keyword, creator type, repository, published date, and modified date. Each page will have a different set of queries for each infographic to display relevant information. We are working to create snapshot visualizations that are suited to each type of data. For example, date dashboards will include a line graph over time and a department specific dashboard may show intradepartmental and

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

interdepartmental collaborations. As we finalize the work here, we'll consider how these dashboards work best for our users and how we might integrate visualizations into the next software release.

Lessons Learned and Continued Challenges

As has been noted in our review of the literature, the dataset repository landscape is new and dispersed, and the metadata describing datasets in these repositories is limited, especially when looking to identify a dataset creator and their affiliation to an institution. Frequently, our team had to work through researcher disambiguation and understanding the researcher's connection to our university as we turned toward large-scale aggregation and harvesting of datasets. While this initially slowed down parts of our work, we ultimately created some viable solutions to identifying our datasets and the work of our researchers. We arrived on a three-pronged strategy for identifying and enhancing metadata for MSU datasets.

First, we looked to survey metadata records for fields that potentially indicated a connection or loose affiliation with Montana State University research. In most cases, our work involved isolating metadata fields that suggested the sources of the dataset. Most of this work was done through manual searches (i.e., a person running searches) to understand how datasets were described and indexed by external data repositories. This work also allowed us to understand coverage of our MSU research and to find the source dataset repositories with the best representation of our research data for the automated work in our next two steps. Second, we query the source repositories for potential matches using the APIs keyword and subject searching functions. We do this by querying each API with several different queries, including "Montana State," "Montana State University," and "MSU." Third, because many data repositories do not log the institutional affiliation of authors, the team looked to identify MSU researchers by going to one of the primary sources of institutional data, the MSU Office of Planning and Analysis (OPA). Most universities will have an institutional data and statistical body that collects and records student enrollment data, faculty numbers, research hires, etc. In our case, we met with OPA to describe our use case for the data and reasons behind the Dataset Search application, and they agreed to provide us with an annual list of names for all tenure-track and non-tenure track faculty. We used this list to query data repository APIs for each individual name. As metadata records were returned, we could use cues from the metadata to attempt to disambiguate the names—for instance, if the researcher was in the Plant Sciences Department at MSU, it was unlikely that they would conduct social science research. Human curators also play an important role in disambiguation.

Even as we started to see success with our strategies for identifying MSU datasets, we also noted a need to build ways to enrich the harvested metadata and to help standardize the metadata. Our API calls were successfully identifying MSU datasets, but the amount and types of metadata returned were sometimes limited and in need of some cleaning up. We saw many of these metadata limitations in

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

the descriptive keywords and subjects for the datasets. We could do much of the manual standardization and cleaning up of records using our administrative interface within MSU Dataset Search. However, we wanted to enhance the subjects and keywords to refine and build out a better level of description. To do this, we harvested keywords from Google Scholar profiles and reconciled those keywords with linked data expressions. In this reconciliation process, we mapped the harvested keywords to Wikidata item entities so that each keyword was associated with a Wikidata URL. We used a Python script to carry out this harvest and reconciliation work; all of our code is openly available in a GitHub repository (Clark et al. 2020). Our working theory was that the keywords and subject terms in Google Scholar profiles were created by the researchers themselves and therefore represent the closest approximation of the type of research they produce and their preferred terms for describing themselves. Adding Wikidata linked data expressions also helps make these enhanced subjects and keywords available to machines to improve indexing via search engines.

Among the other lessons learned and challenges faced, the team needed to understand what a successful index of our datasets looked like. Would 80% of our dataset output provide enough scope and a working inventory of data production at our institution? The completeness of the index was a quantifiable element that we needed to reconcile. We ultimately understand that our index likely won't be a comprehensive list of datasets from MSU researchers. False positives are common when querying data repositories for full names, and we also anticipate that we are not finding all datasets that have been published by MSU researchers. In the future, we can help reduce false positives and increase completeness by integrating ORCID with our tool, and by using CrossRef and DataCite DOI metadata to connect datasets with any associated publications that include institutional affiliation.

Dataset Search should also be findable in external environments like commercial search engines and Google Dataset Search. We noted above how our data model was predicated on metadata fields for discovery settings, like commercial search engines, and how this focus forced us to modify the data model to accommodate technical and administrative metadata. This was one of our first lessons learned, but there were other solutions that became part of this work. The team pursued what we started to call "architectures for findability" which led to particular patterns of markup for our datasets. We wanted to allow for machine processes and intelligent software agents to discover and understand our datasets and we wanted our datasets to be indexed in Google Dataset Search. Working backward from these goals, we adhered to the best practices for dataset markup released and supported by Schema.org (Google Developers 2020b). In its simplest form, we included the dataset markup on individual dataset items as part of the HTML webpages. We also built an XML sitemap (Google Support 2020) that listed and identified our structured data markup for web indexing tools[2]. We continue to

3 For those interested in learning more about Dataset structured data markup, the Google Developers site offers a helpful explanation and steps for Dataset markup implementation (Google Developers 2020a).

JeSLIB 2021; 10(1): e1189
https://doi.org/10.7191/jeslib.2021.1189

monitor the success and return on these markup activities using validation tools like the Rich Results Testing Tool (Google 2020c), and analytics tools like Google Search Console (Google 2020b) to confirm correct markup patterns and understand the coverage and indexing rates of our datasets in search engines. Benchmarking the appearance of our dataset item records in repositories like Google Dataset Search and DataCite will provide additional insight here. This is a work in progress, but we have seen results for these markup activities in other library properties. A similar markup and indexing project for our library databases (Clark and Rossmann 2017) guides our work here. In that research, we saw increased traffic and organic search referrals based on markup and optimized search engine indexing routines. We are following the same model here and expect to see a similar increased visibility for our datasets.

And finally, our team wanted to find ways to streamline the user interface for dataset retrieval. We worked with user interface designer Lorraine Chuen (Chuen 2020) to create streamlined interfaces that are beautiful and usable. Chuen also helped conduct an expert review of the tool to streamline patterns of use and improve users' navigation through the system. Among the highlights of this work: a clean, simple design using MSU's institutional branding; improved scannability of the page by changing the layout and adding a "Details" panel for metadata; an "Access Dataset" button that clearly guides away from our search interface to the data repository where they can access the dataset; and removal of administrative screens from public view. Chuen's design and expert review have led to a much improved interface.

Future Directions

We see several future directions for Dataset Search. First, we have not conducted large-scale user testing or other assessment of the tool. Next steps could include continuing to monitor our search engine optimization protocols to ensure that the tool is discoverable on the web; conducting user testing locally and updating the user interface in response to any remaining usability issues; installing Google Analytics to understand user traffic; and adding a contact form to the site to support direct user feedback.

After completing the pilot project providing discovery for restricted data records with the Center for American Indian and Rural Health Equity (described above), we may reach out to other research centers who would benefit from increased transparency by sharing metadata records for restricted data. Dataset Search metadata records could also support discovery of data from in-progress projects that are stored locally at MSU, thus encouraging new collaborations and accelerating scientific discoveries.

With our structured data activities and enhancements, we are also noting some new possibilities around sharing the datasets and reuse of the data. MSU Dataset Search has a default API that is under development, but it is not standardized or documented. The team recognizes that there is some useful work to be completed

here and has begun looking at new API formats that could benefit the data community if implemented. A member of our team has been working with the Research Object Crate (RO-Crate) standards group to shape the emerging standard for use with datasets and to pilot a use case of RO-Crate. RO-Crate is "lightweight approach to packaging research data with their structured metadata, rephrasing the Research Object model as Schema.org annotations to formalize a JSON-LD format that can be used independently of infrastructure" (Carragáin et al. 2019). More specifically, our team is looking to standardize the Dataset Search API using the RO-Crate standard which would allow us to connect our API implementation to the broader work of the research objects community and help shape documentation and use of our API.

Dataset Search is built with open source code (Clark et al. 2020) and we have outlined a straightforward installation process; the front-end design is also customizable to match the branding of any institution. We therefore hope that the Dataset Search will be adopted by other small- and mid-sized institutions who are looking for a lightweight tool to promote discovery and access for their local research data. As a member of the Data Discovery Collaboration (DDC 2020), the Dataset Search project benefits from alignment with other similar projects, and we will continue to pursue connections with the data discovery community and explore how the functionalities of the Dataset Search tool can be integrated with other data catalog infrastructures such as the NYU-developed Data Catalog software (Lamb and Larson 2016). Our automatic harvesting routine could also be integrated with data repository software such as Dataverse (Dataverse 2020).

Conclusion

As research data sharing grows, institutions are increasingly building initiatives that support discovery, access, and reuse for published data. Montana State University's Dataset Search is designed as a lightweight, open-source solution that supports discovery and reporting for research data created by researchers at our institution. The Dataset Search application provides five core features to support dataset discovery: a streamlined browse and search interface, a data model based on dataset discovery, a harvesting process for finding and vetting datasets stored in external repositories, an administrative interface for managing the creation, ingest, and maintenance of dataset records, and a dataset visualization interface to demonstrate how data is produced and used by MSU researchers. Dataset Search is designed to be easily customized and implemented by other institutions to improve search and discovery and therefore amplify the impact and benefits of research data.

Acknowledgments

This project is made possible in part by the National Network of Libraries of Medicine, Pacific Northwest Region. The mission of the National Network of Libraries of Medicine (NNLM) is to advance the progress of medicine and improve the public health by providing all U.S. health professionals with equal access to

biomedical information and improving the public's access to information to enable them to make informed decisions about their health. To learn more, visit https://nnlm.gov.

This project is made possible in part by the Institute of Museum and Library Services, through grant #LG-89-18-0225-18. The IMLS is the primary source of federal support for the nation's libraries and museums. They advance, support, and empower America's museums, libraries, and related organizations through grantmaking, research, and policy development. Their vision is a nation where museums and libraries work together to transform the lives of individuals and communities. To learn more, visit https://www.imls.gov.

Many thanks to Dataset Search's 2018-2019 Advisory Board: Jeffrey Grethe, Kevin Read, Nicole Contaxis, and Cynthia Hudson-Vitale.

Montana State University Dataset Search is a member of the Data Discovery Collaboration.

Disclosures

The substance of this article is based upon a lightning talk at RDAP Summit 2020: "Dataset Search: A lightweight, community-built tool to support research data discovery at small and mid-sized institutions" available at https://osf.io/qctfa.

Data Availability

Code associated with this paper is available in Zenodo, via Github:

Clark, Jason A., James Espeland, Jakob Schultz, Kyle Hagerman, Daniel Laden, and Sara Mannheimer. 2020. "msulibrary/dataset-Search: Prototype release (v0.9)." Zenodo. https://doi.org/10.5281/zenodo.4046567.

MSU Dataset Search is available at:
https://arc.lib.montana.edu/msu-dataset-search.

References

Arlitsch, Kenning, and Patrick S. O'Brien. 2012. "Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar." *Library Hi Tech* 30(1): 60–81. https://doi.org/10.1108/07378831211213210

Baker, Edward N. 2017. "Data Archiving and Availability in an Era of Open Science." *IUCrJ* 4(Pt 1): 1–2. https://doi.org/10.1107/S2052252516020340

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2013. "GenBank." *Nucleic Acids Research* 41(Database issue): D36-42. https://doi.org/10.1093/nar/gks1195

Boulton, Geoffrey, Philip Campbell, Brian Collins, Peter Elias, Wendy Hall, Graeme Laurie, Onora O'Neill, et al. 2012. Science as an Open Enterprise. London, UK: The Royal Society—Science Policy Centre.

CAIRHE. 2020. "Center for American Indian and Rural Health Equity — Montana State University." https://web.archive.org/web/20200512081726/http://www.montana.edu/cairhe

Carragáin, Eoghan Ó, Carole Goble, Peter Sefton, and Stian Soiland-Reyes. 2019. "A Lightweight Approach to Research Object Data Packaging." In *Proceedings of the Bioinformatics Open Source Conference (BOSC)*, ISMB/ECCB 2019 (Session 1, Part 1075). Basel, Switzerland: Zenodo. https://doi.org/10.5281/ZENODO.3250687

Center for Qualitative and Multi-Method Inquiry. 2020. "Qualitative Data Repository." https://qdr.syr.edu

CERN Data Centre. 2020. "Zenodo." https://web.archive.org/web/20200524175824/https://zenodo.org

Chen, Xiaoling, Anupama E. Gururaj, Burak Ozyurt, Ruiling Liu, Ergin Soysal, Trevor Cohen, Firat Tiryaki, et al. 2018. "DataMed – an Open Source Discovery Index for Finding Biomedical Datasets." *Journal of the American Medical Informatics Association* 25(3): 300–308. https://doi.org/10.1093/jamia/ocx121

Chuen, Lorraine. 2020. Professional website. http://web.archive.org/web/20201027195026/http://lorrainechuen.com

Clark, Jason A., James Espeland, Jakob Schultz, Kyle Hagerman, Daniel Laden, and Sara Mannheimer. 2020. "msulibrary/dataset-Search: Prototype release (v0.9)." Zenodo. https://doi.org/10.5281/zenodo.4046567

Clark, Jason A., and Doralyn Rossmann. 2017. "The Open SESMO (Search Engine & Social Media Optimization) Project: Linked and Structured Data for Library Subscription Databases to Enable Web-Scale Discovery in Search Engines." *Journal of Web Librarianship* 11(3–4): 172–193. https://doi.org/10.1080/19322909.2017.1378148

CoreTrustSeal. 2020. "Core Trustworthy Data Repositories Requirements." https://web.archive.org/web/20200408004456/https://www.coretrustseal.org/why-certification/requirements

Dataverse. 2020. "The Global Dataverse Community Consortium." http://dataversecommunity.global/home

DDC. 2020. "The Data Discovery Collaboration." https://web.archive.org/web/20200525191325/https://www.datacatalogcollaborationproject.org

Digital Science & Research Ltd. 2020. "Figshare." https://web.archive.org/web/20200523121005/https://figshare.com

Dryad. 2020. "Dryad Digital Repository." https://web.archive.org/web/20200524165914/https://datadryad.org/stash

Dryad Digital Repository. 2011. "Joint Data Archiving Policy." http://wiki.datadryad.org/Joint_Data_Archiving_Policy_(JDAP)

Federer, Lisa M., Christopher W. Belter, Douglas J. Joubert, Alicia Livinski, Ya-Ling Lu, Lissa N. Snyders, and Holly Thompson. 2018. "Data Sharing in PLOS ONE: An Analysis of Data Availability Statements." Edited by Jelte M. Wicherts. *PLOS ONE* 13(5): e0194768. https://doi.org/10.1371/journal.pone.0194768

Fenner, Martin, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, et al. 2019. "A Data Citation Roadmap for Scholarly Data Repositories." *Scientific Data* 6(1): 28. https://doi.org/10.1038/s41597-019-0031-8

Fienberg, Stephen E., Margaret E. Martin, National Research Council (U S.) Committee on National Statistics, and National Research Council (U S.) Commission on Behavioral and Social Sciences and Education. 1985. *Sharing Research Data*. National Academies.

Google. 2020a. "Google Scholar." https://scholar.google.com

———. 2020b. "Google Search Console." http://web.archive.org/web/20200920134707/https://search.google.com/search-console/about

———. 2020c. "Rich Results Testing Tool." http://web.archive.org/web/20200920070004/https://search.google.com/test/rich-results

Google Developers. 2020a. "Dataset—How to Add Structured Data." http://web.archive.org/web/20200920070511/https://developers.google.com/search/docs/data-types/dataset#add-structured-data

———. 2020b. "Google Search for Developers Reference - Dataset." https://web.archive.org/web/20200529145106/https://developers.google.com/search/docs/data-types/dataset

Google Support. 2020. "Google Search Console Help - Build and Submit a Sitemap." https://web.archive.org/web/20200514184213/https://support.google.com/webmasters/answer/183668

Haaker, Maureen, and Bethany Morgan-Brett. 2017. "Developing Research-Led Teaching: Two Cases of Practical Data Reuse in the Classroom." *SAGE Open* 7(2): 2158244017701800. https://doi.org/10.1177/2158244017701800

Harvard College. 2020. "Harvard Dataverse." https://dataverse.harvard.edu

Holdren, John P. 2013. "Increasing Access to the Results of Federally Funded Scientific Research." White House Office of Science and Technology Policy, February. https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

Hudson-Vitale, Cynthia R., Richard P. Johnson, Judy Ruttenberg, and Jeffrey R. Spies. 2017. "SHARE: Community-Focused Infrastructure and a Public Goods, Scholarly Database to Advance Access to Research." *D-Lib Magazine* 23(5/6). https://doi.org/10.1045/may2017-vitale

Kim, Youngseek, and Jeffrey M. Stanton. 2016. "Institutional and Individual Factors Affecting Scientists' Data-Sharing Behaviors: A Multilevel Analysis." *Journal of the Association for Information Science and Technology* 67(4): 776–799. https://doi.org/10.1002/asi.23424

Kim, Youngseek, and Ping Zhang. 2015. "Understanding Data Sharing Behaviors of STEM Researchers: The Roles of Attitudes, Norms, and Data Repositories." *Library & Information Science Research* 37(3): 189–200. https://doi.org/10.1016/j.lisr.2015.04.006

Kriesberg, Adam, Rebecca D. Frank, Ixchel M. Faniel, and Elizabeth Yakel. 2013. "The Role of Data Reuse in the Apprenticeship Process." *Proceedings of the American Society for Information Science and Technology* 50(1): 1–10. https://doi.org/10.1002/meet.14505001051

Lamb, Ian, and Catherine Larson. 2016. "Shining a Light on Scientific Data: Building a Data Catalog to Foster Data Sharing and Reuse." *The Code4Lib Journal* 32(April). http://journal.code4lib.org/articles/11421

Lin, Dawei, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, et al. 2020. "The TRUST Principles for Digital Repositories." *Scientific Data* 7(1): 144. https://doi.org/10.1038/s41597-020-0486-7

Mannheimer, Clark, Espeland, and Hagerman. 2019. "Building a Dataset Search for Institutions: Project Update." *Publications* 7(2): 29. https://doi.org/10.3390/publications7020029

Mannheimer, Sara, and J. A. Clark. 2019. "Dataset Search: A Lightweight Tool to Promote Discovery of Health Sciences Research Data - National Network of Libraries of Medicine, Pacific Northwest Region (NNLM PNR) Research & Data Engagement Award." http://web.archive.org/web/20200525190653/https://saramannheimer.com/wp-content/uploads/2020/05/Mannheimer_2019_NNLM_PNR_Proposal.pdf

Mannheimer, Sara, Leila Sterman, and Susan Borda. 2016. "Discovery and Reuse of Open Datasets: An Exploratory Study." *Journal of EScience Librarianship* 5(1): e1091. https://doi.org/10.7191/jeslib.2016.1091

Marcial, Laura Haak, and Bradley M. Hemminger. 2010. "Scientific Data Repositories on the Web: An Initial Survey." *Journal of the American Society for Information Science and Technology* 61(10): 2029–48. https://doi.org/10.1002/asi.21339

Montana State University. 2018. "A Prototype for an Institutional Research Data Index. Funded by the Institute of Museum and Library Services LG-89-18-0225-18." https://www.imls.gov/grants/awarded/lg-89-18-0225-18

———. 2020. "Quick Facts: 2019-2020 - Office of Planning & Analysis." https://web.archive.org/web/20200525174939/http://www.montana.edu/opa/facts/quick.html

MSU Dataset Search. 2020. https://arc.lib.montana.edu/msu-dataset-search

MSU Library. 2020. "People in the Library." https://web.archive.org/web/20200525174829/https://www.lib.montana.edu/people/search.php?staff=all

National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. Washington, D.C.: National Academies Press. https://doi.org/10.17226/25303

National Institutes of Health. 2003. "NIH Data Sharing Policy and Implementation Guidance." https://web.archive.org/web/20200502190309/https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Science Foundation. 2011. "Dissemination and Sharing of Research Results." http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nature Biotechnology. 2009. "Credit Where Credit Is Overdue." *Nature Biotechnology* 27(7): 579–579. https://doi.org/10.1038/nbt0709-579

Noy, Natasha. 2020. "Discovering Millions of Datasets on the Web." Google: The Keyword. January 23, 2020. https://web.archive.org/web/20200515113720/https://www.blog.google/products/search/discovering-millions-datasets-web

NSIDC. 2020. "National Snow and Ice Data Center." https://web.archive.org/web/20200524071213/https://nsidc.org

Pampel, Heinz, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf. 2013. "Making Research Data Repositories Visible: The Re3data.Org Registry." *PLOS ONE* 8(11): e78080. https://doi.org/10.1371/journal.pone.0078080

Pasquetto, Irene V., Christine L. Borgman, and Morgan F. Wofford. 2019. "Uses and Reuses of Scientific Data: The Data Creators' Advantage." *Harvard Data Science Review* 1(2). https://doi.org/10.1162/99608f92.fc14bf2d

Piwowar, Heather A., and Todd J. Vision. 2013. "Data Reuse and the Open Data Citation Advantage." *PeerJ* 1(October): e175. https://doi.org/10.7717/peerj.175

PLOS. 2014. "Data Availability." http://journals.plos.org/plosone/s/data-availability

Poole, Alex H. 2015. "How Has Your Science Data Grown? Digital Curation and the Human Factor: A Critical Literature Review." *Archival Science* 15(2): 101–139. https://doi.org/10.1007/s10502-014-9236-y

Pronk, Tessa E. 2019. "The Time Efficiency Gain in Sharing and Reuse of Research Data." *Data Science Journal* 18(1): 10. https://doi.org/10.5334/dsj-2019-010

Read, Kevin B, Nicole Contaxis, Ian Lamb, Catherine Larson, and Alisa Surkis. 2018. "A Cross-Institutional Data Discovery Collaboration: Indexing Institutional Research Data." In *Proceedings of the AMIA Informatics Summit*. San Francisco, CA. https://goo.gl/3diLu7

Registry of Research Data Repositories (Re3data.Org). 2020. "Data Repositories Filtered by Country: United States of America." https://www.re3data.org/search?query=&countries%5B%5D=USA

Sansone, Susanna-Assunta, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, George Alter, Jeffrey S. Grethe, Hua Xu, Ian M. Fore, et al. 2017. "DATS, the Data Tag Suite to Enable Discoverability of Datasets." *Scientific Data* 4(June): 170059. https://doi.org/10.1038/sdata.2017.59

Scerri, Antony, John Kuriakose, Amit Ajit Deshmane, Mark Stanger, Peter Cotroneo, Rebekah Moore, Raj Naik, and Anita de Waard. 2017. "Elsevier's Approach to the BioCADDIE 2016 Dataset Retrieval Challenge." *Database* 2017(January). https://doi.org/10.1093/database/bax056

Schema.org. 2020a. "Data and Datasets." https://schema.org/docs/data-and-datasets.html

———. 2020b. "MeasurementTechnique - Schema.Org Property." http://web.archive.org/web/20200508100048/https://schema.org/measurementTechnique

———. 2020c. "VariableMeasured - Schema.Org Property." http://web.archive.org/web/20200506232236/https://schema.org/variableMeasured

Shen, Yi. 2017. "Burgeoning Data Repository Systems, Characteristics, and Development Strategies: Insights of Natural Resources and Environmental Scientists." *Data and Information Management* 1(2): 115–123. https://doi.org/10.1515/dim-2017-0009

Sterman, Leila, and Jason Clark. 2017. "Citations as Data: Harvesting the Scholarly Record of Your University to Enrich Institutional Knowledge and Support Research." *College & Research Libraries* 78(7). https://doi.org/10.5860/crl.78.7.952

Taichman, Darren B., Peush Sahni, Anja Pinborg, Larry Peiperl, Christine Laine, Astrid James, Sung-Tae Hong, et al. 2017. "Data Sharing Statements for Clinical Trials - A Requirement of the International Committee of Medical Journal Editors." *The New England Journal of Medicine* 376(23): 2277–2279. https://doi.org/10.1056/NEJMe1705439

Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. 2015. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide." *PLOS ONE* 10(8): e0134826. https://doi.org/10.1371/journal.pone.0134826

University of Michigan. 2020. "Deep Blue Data."
https://web.archive.org/web/20200511095414/https://deepblue.lib.umich.edu/data

Wallis, Jillian C., Elizabeth Rolando, and Christine L. Borgman. 2013. "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology." *PLOS ONE* 8(7): e67332. https://doi.org/10.1371/journal.pone.0067332

Wikidata. 2020. "Wikidata Main Page."
http://web.archive.org/web/20200919221815/https://www.wikidata.org/wiki/Wikidata:Main_Page

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3(March): 160018. https://doi.org/10.1038/sdata.2016.18

Witt, Michael. 2008. "Institutional Repositories and Research Data Curation in a Distributed Environment." *Library Trends* 57(2): 191–201. https://doi.org/10.1353/lib.0.0029