publications

Case Report

Building a Dataset Search for Institutions: Project Update

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

Montana State University Library, Bozeman, Montana 59717, USA; jaclark@montana.edu (J.A.C.); james.espeland@montana.edu (J.E.); kyle.hagerman@student.montana.edu (K.H.)
* Correspondence: sara.mannheimer@montana.edu

Received: 13 February 2019; Accepted: 22 April 2019; Published: 24 April 2019

Abstract: Most out-of-the-box institutional repository systems do not provide the workflows and metadata features required for research data. Consequently, many libraries now support two institutional repository systems—one for publications, and one for research data—even when there are nearly a thousand data repositories in the United States, many of which provide services and policies that ensure their trustworthiness and suitability for research data. Libraries are either increasing spending by purchasing data repository solutions from vendors, or replicating work by building, customizing, and managing individual instances of data repository software. This article gives an update on a potential solution to this issue: An in-progress prototype for an open source Dataset Search tool that promotes discovery and reuse of institutional research datasets through automatic metadata harvesting and search engine optimization. Once finished, the Dataset Search tool has the potential to support three key impacts: Increasing discovery, reuse, and citation of research data; reinforcing the idea that research data are a legitimate scholarly product; and promoting community-owned systems that require less resource expenditure.

Keywords: research data; dataset; search; index; data discovery; data reuse

1. Introduction and Background

Academic libraries are increasingly participating in research data publication and preservation [1–3]. However, out-of-the-box institutional repository (IR) systems like DSpace [4] and Digital Commons [5] 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 are often published in a preliminary state, then updated with new versions as projects progress [6] (p. 52). 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.

Some repository systems are designed specifically for research data, but they are resource-intensive. Open source systems like Dataverse [7], CKAN [8], DKAN [9], and Samvera [10] require developer hours and data storage infrastructure. Vendor solutions like Figshare for Institutions [11] and Tind [12] 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. In addition, data repository systems require that libraries support permanent storage for the datasets stored within, and research data pose unique digital preservation challenges, including heterogeneous file types and very large file sizes [13]. As of 2018, there are nearly a thousand data repositories in the United States [14], many of which provide services and policies that ensure their trustworthiness and suitability for research data. We suggest that small and mid-sized institutions can both conserve their limited resources and increase the discovery of institutional research datasets by directing their researchers to one of these third-party repositories, and then providing local access to the published datasets through a searchable and discoverable index.

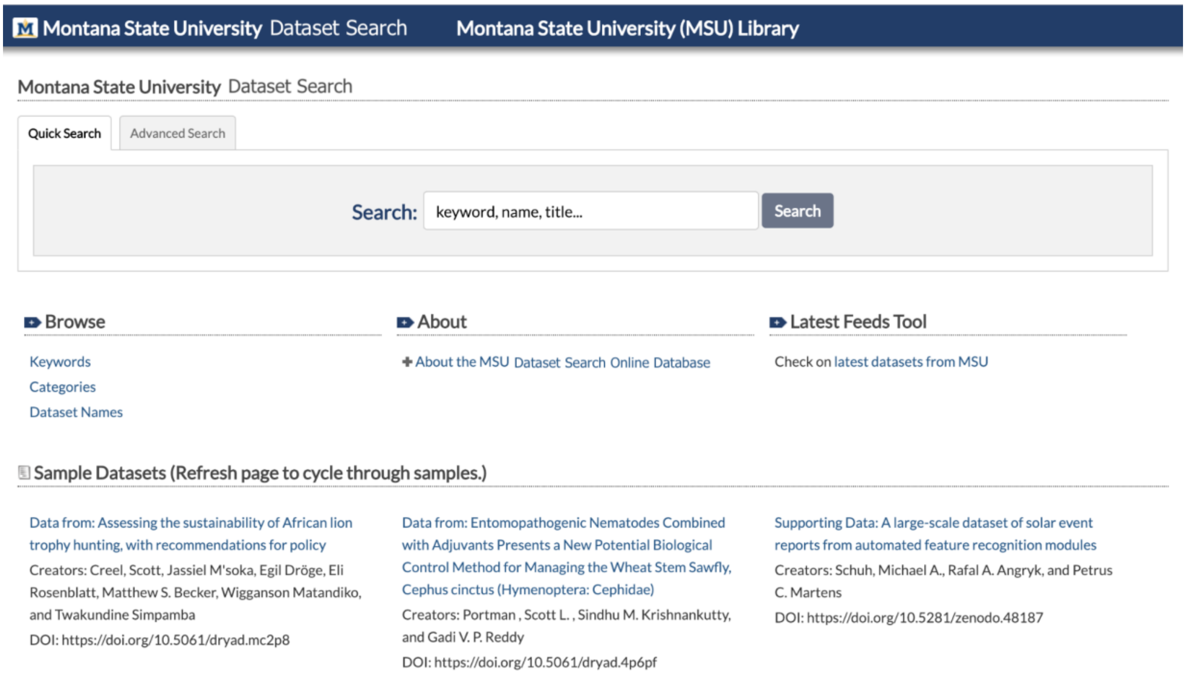
2. Dataset Search

This article describes an in-progress project that could provide a solution to the issues outlined above: An open source, scalable, sustainable, and standardized Dataset Search tool that will promote discovery and reuse of research datasets while expending fewer resources than those required for an institutional data repository. Unlike a data repository, the in-progress prototype for the Montana State University (MSU) Dataset Search[1] 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. 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, Dataset Search does not store actual datasets, but rather provides metadata so that visitors can find the repositories where the datasets are stored. Dataset Search builds on similar projects such as the Data Catalog Collaboration Project [16,17], NIH DataMed [18], and SHARE [19], adding three innovations. 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 promotes discovery through leading commercial web search engines. Third, Dataset Search will automatically generate new descriptive metadata for individual datasets using external topic mining of scholarly profile sources like ORCID and Google Scholar Profiles.

3. Implementation

The Dataset Search tool is based on an existing system at MSU Library that harvests citation information for published journal articles [20]. Dataset Search automatically harvests metadata for MSU-affiliated research datasets using data repository feeds and APIs, which are parsed using PHP scripts. After undergoing deduplication and human curatorial review, metadata records are saved in a local database. A user interface allows users to search and access the metadata records (see Figure 1). Once the Dataset Search tool is finished, code will be made available in Github [21] and the project will be shared more widely with the community. Dataset Search is funded by an Institute of Museum and Library Services (IMLS) National Leadership Grant [22], with a funding period of one year. As shown in Figure 2, the project began in October 2018, and will continue until September 2019. As of the writing of this article, the project is approximately at its mid-point.

[1] Dataset Search was formerly named the Institutional Research Data Index. This work was presented at the Open Repositories 2018 Conference in the presentation "A Prototype for the Institutional Research Data Index" [15].



Montana State University Dataset Search
Montana State University (MSU) Library
Montana State University Dataset Search
Quick Search
Advanced Search
Search:
keyword, name, title...
Search
Browse
Keywords
Categories
Dataset Names
About
About the MSU Dataset Search Online Database
Latest Feeds Tool

Check on latest datasets from MSU

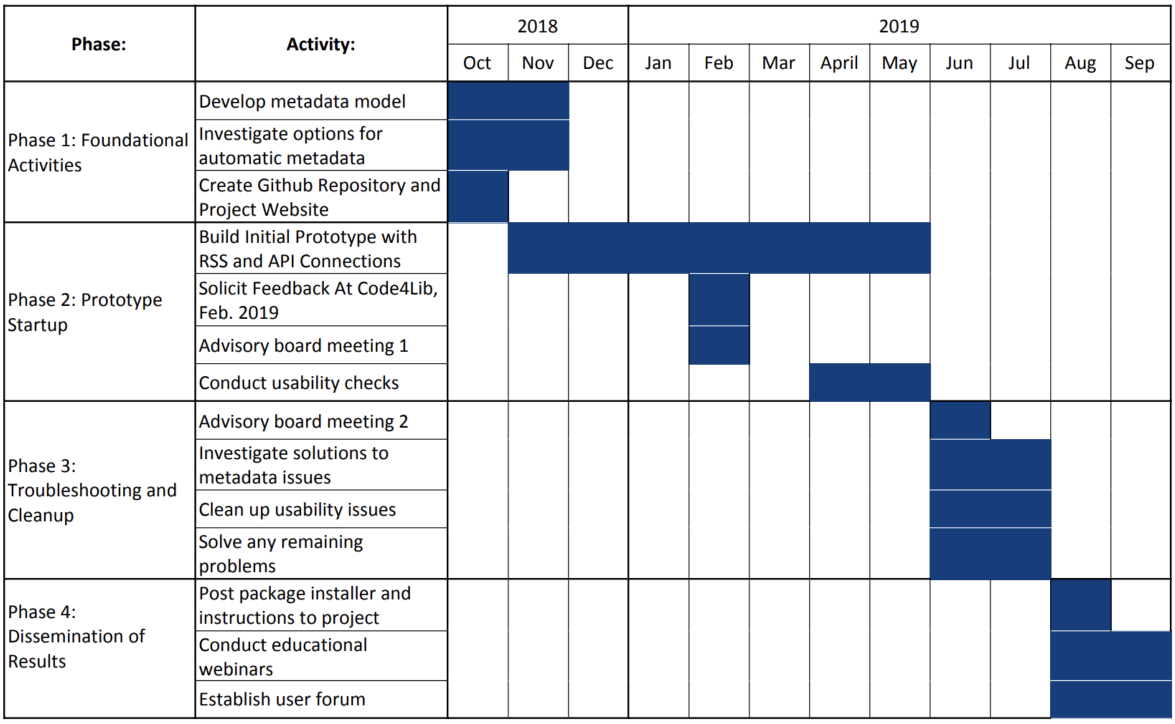
Sample Datasets (Refresh page to cycle through samples.)

Data from: Assessing the sustainability of African lion trophy hunting, with recommendations for policy
Creators: Creel, Scott, Jassiel M'soka, Egil Dröge, Eli Rosenblatt, Matthew S. Becker, Wigganson Matandiko, and Twakundine Simpamba
DOI: https://doi.org/10.5061/dryad.mc2p8

Data from: Entomopathogenic Nematodes Combined with Adjuvants Presents a New Potential Biological Control Method for Managing the Wheat Stem Sawfly, Cephus cinctus (Hymenoptera: Cephidae)
Creators: Portman , Scott L. , Sindhu M. Krishnankutty, and Gadi V. P. Reddy
DOI: https://doi.org/10.5061/dryad.4p6pf

Supporting Data: A large-scale dataset of solar event reports from automated feature recognition modules
Creators: Schuh, Michael A., Rafal A. Angryk, and Petrus C. Martens
DOI: https://doi.org/10.5281/zenodo.48187

Figure 1 The Dataset Search prototype.



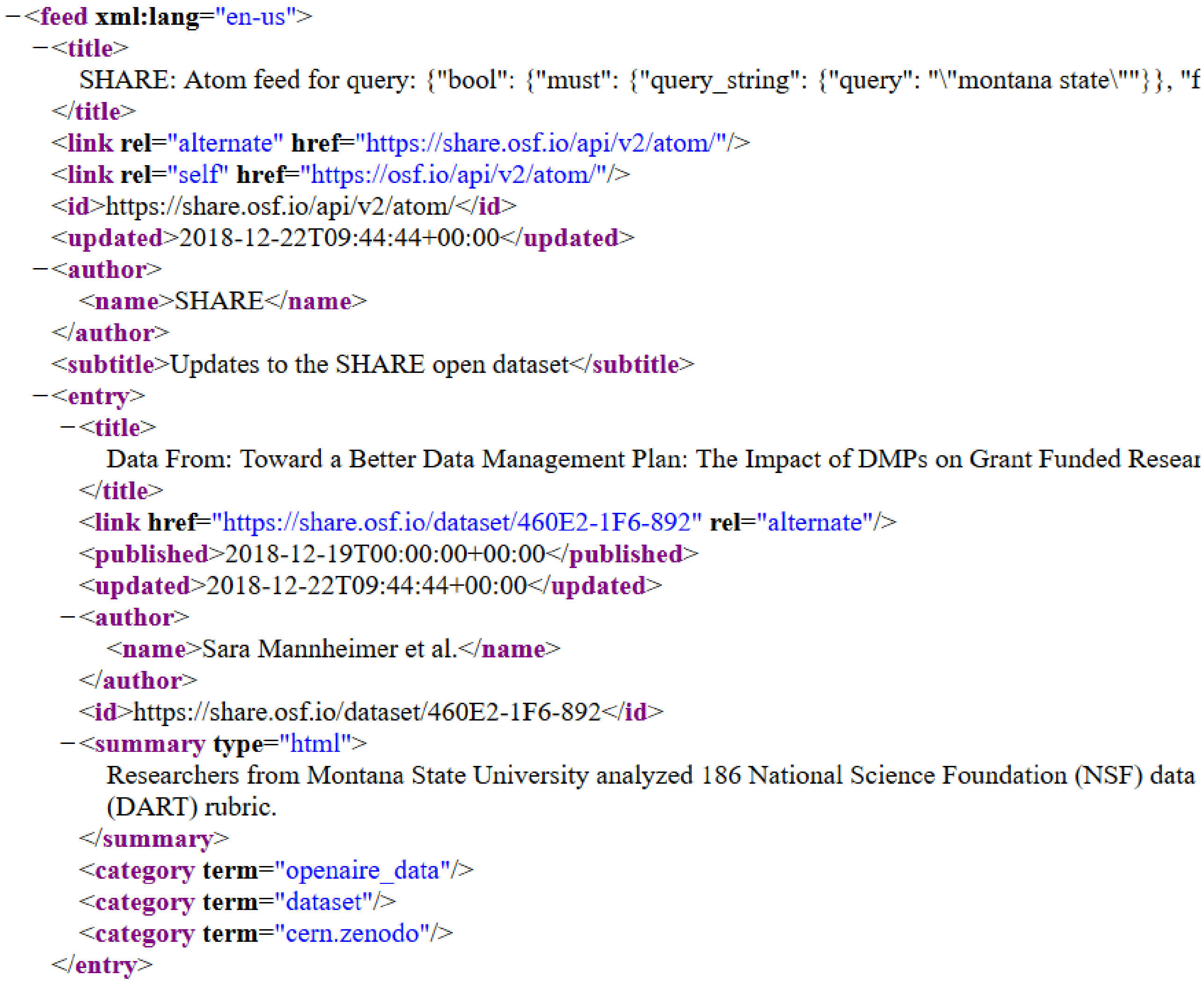
Phase:	Activity:	2018			2019								
		Oct	Nov	Dec	Jan	Feb	Mar	April	May	Jun	Jul	Aug	Sep
Phase 1: Foundational Activities	Develop metadata model	■	■										
	Investigate options for automatic metadata	■	■										
	Create Github Repository and Project Website	■											
Phase 2: Prototype Startup	Build Initial Prototype with RSS and API Connections		■	■	■	■	■	■	■				
	Solicit Feedback At Code4Lib, Feb. 2019					■							
	Advisory board meeting 1					■							
	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											■	
	Conduct educational webinars											■	■
	Establish user forum											■	■

Figure 2. The timeline for the Dataset Search project.

3.1. Data Repository Automated Feed Harvesting

As of this halfway point in the project, we have created a workflow that programmatically harvests dataset metadata using feed data from various data repositories. As the project continues to progress, a similar workflow will be developed for APIs. For feed data, a list of feed URLs is stored in a "feeds" table in the database and AJAX calls are used to fetch the contents of each feed upon demand. Since the structure of the feed data varies from one feed to another, the Dataset Search tool uses an XML file that contains a map of the XML tag structure of each unique feed to guide it through the harvesting of metadata. This XML file informs the Dataset Search tool of the location within the feed of seven key metadata fields: Repository name, title, description, author, publication date, link (web link to the dataset), and UID (universal ID such as DOI). These seven fields are not mandatory, and metadata records will still be created if some of these fields are not present. Human curators will review each record and fill in missing information if it is available. We are also experimenting with some autogeneration of metadata to help with missing information which we take up in more detail in Section 3.4.

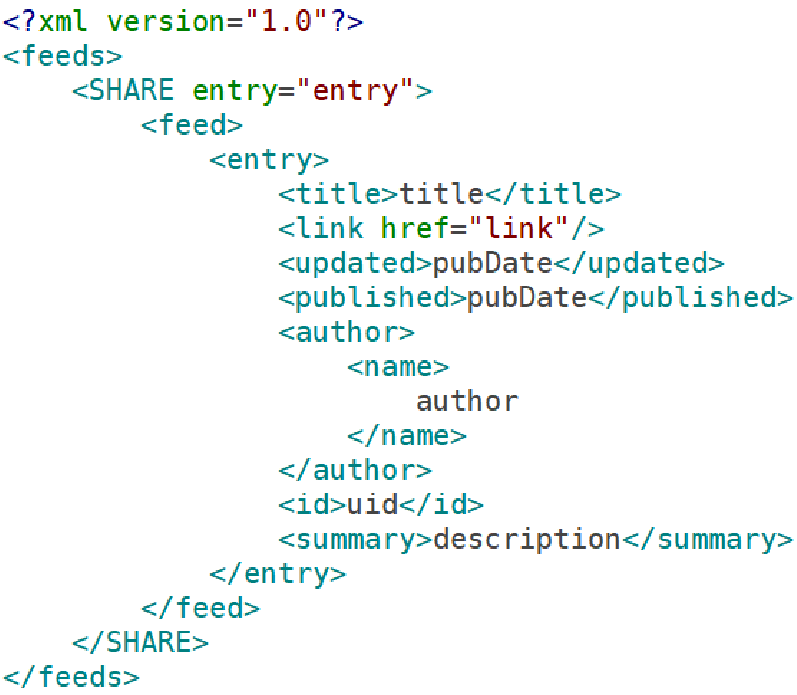
By way of example, consider the excerpt from a SHARE atom feed shown in Figure 3.



```
–<feed xml:lang="en-us">
  –<title>
      SHARE: Atom feed for query: {"bool": {"must": {"query_string": {"query": "\"montana state\""}}, "f
    </title>
    <link rel="alternate" href="https://share.osf.io/api/v2/atom/"/>
    <link rel="self" href="https://osf.io/api/v2/atom/"/>
    <id>https://share.osf.io/api/v2/atom/</id>
    <updated>2018-12-22T09:44:44+00:00</updated>
  –<author>
      <name>SHARE</name>
    </author>
    <subtitle>Updates to the SHARE open dataset</subtitle>
  –<entry>
    –<title>
        Data From: Toward a Better Data Management Plan: The Impact of DMPs on Grant Funded Resear
      </title>
      <link href="https://share.osf.io/dataset/460E2-1F6-892" rel="alternate"/>
      <published>2018-12-19T00:00:00+00:00</published>
      <updated>2018-12-22T09:44:44+00:00</updated>
    –<author>
        <name>Sara Mannheimer et al.</name>
      </author>
      <id>https://share.osf.io/dataset/460E2-1F6-892</id>
    –<summary type="html">
        Researchers from Montana State University analyzed 186 National Science Foundation (NSF) data
        (DART) rubric.
      </summary>
      <category term="openaire_data"/>
      <category term="dataset"/>
      <category term="cern.zenodo"/>
    </entry>
```

Figure 3. An excerpt of SHARE Atom Feed.

The feed structure shown in Figure 4, below, maps the locations of the desired metadata elements in the SHARE feed. The Dataset Search tool uses the feed structure XML map to locate each metadata element in the SHARE feed and extract this data into a database record. Before the record is entered into the database, a deduplication process will be performed to ensure that each dataset is entered into the database only once. Since DOIs are unique identifiers, presence of DOI metadata could be used to filter out duplicates. In the absence of DOIs, a dataset title could be used to assist with duplicate detection. Normalization of the titles might be necessary to account for differences involving capitalization, use of 'and' versus '&', and use of non-alphanumeric characters. The team has not yet finalized the deduplication process, but has plans to finalize the process by August 2019.

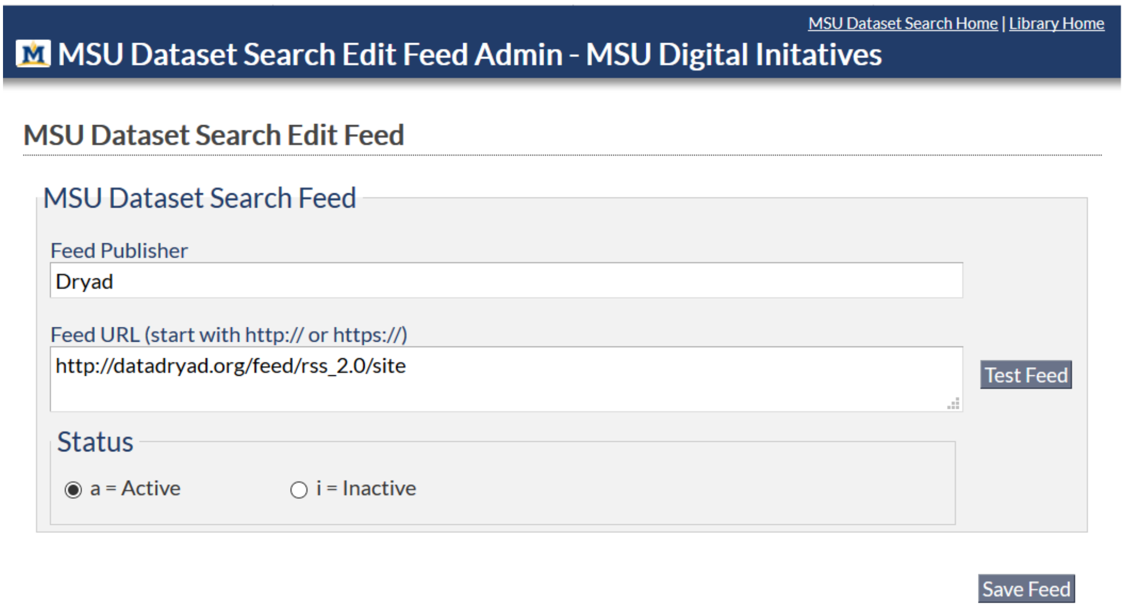


```
<?xml version="1.0"?>
<feeds>
    <SHARE entry="entry">
        <feed>
            <entry>
                <title>title</title>
                <link href="link"/>
                <updated>pubDate</updated>
                <published>pubDate</published>
                <author>
                    <name>
                        author
                    </name>
                </author>
                <id>uid</id>
                <summary>description</summary>
            </entry>
        </feed>
    </SHARE>
</feeds>
```

Figure 4. feedStructure.xml.

When adding a new data repository for metadata harvest, a new feed structure must be generated. The Dataset Search system will provide a guided interface to lead the user through this process. Once a feed has been introduced to the system, PHP scripts will parse out the relevant tags that contain desired metadata fields. When prompted, users can visually identify these fields, building the feed structure automatically.

Users wishing to adopt the Dataset Search open source code will first add a selected feed using the Edit Feed web form, as shown in Figure 5.



MSU Dataset Search Edit Feed Admin - MSU Digital Initatives

MSU Dataset Search Edit Feed

MSU Dataset Search Feed

Feed Publisher
Dryad

Feed URL (start with http:// or https://)
http://datadryad.org/feed/rss_2.0/site

Status

● a = Active ○ i = Inactive

Figure 5. The Edit Feed web form.

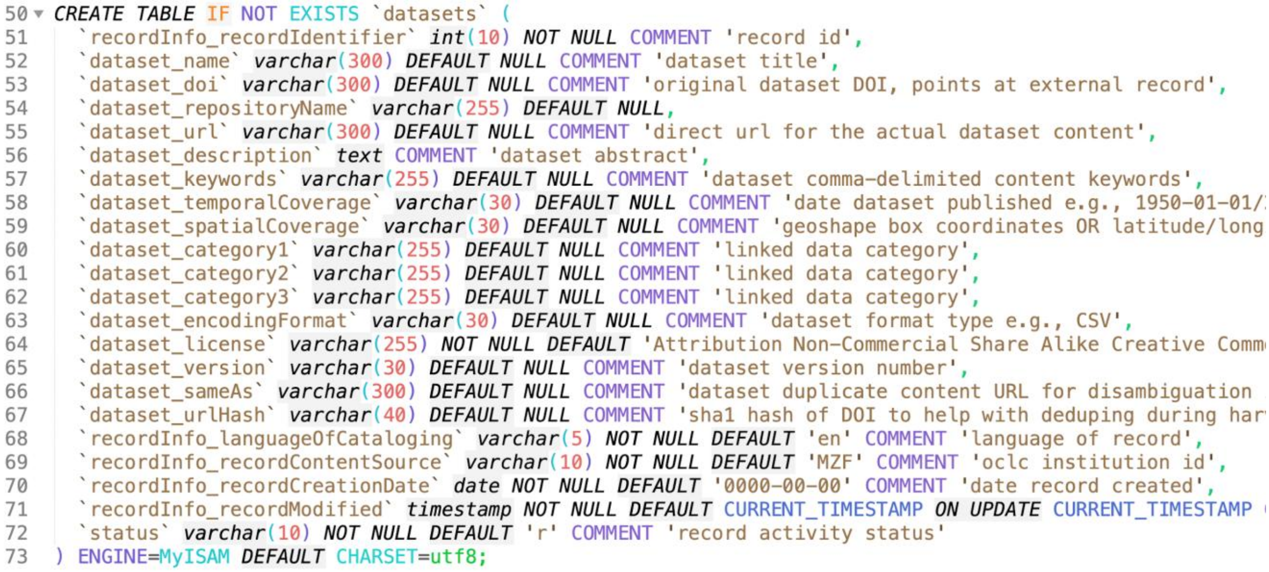
Users will then use the guided interface to add an entry to the XML feed structure file (see sample of XML feed structure in Figure 4). The guided interface is a web form that will assist a user in identification of relevant feed tags and creation of the new entry for the XML feed structure file.

3.2. Repository Selection

It would be impossible for the Dataset Search tool to harvest the nearly one thousand data repositories in the United States. The project team will initially focus on harvesting other data repository aggregation sites such as DataCite, SHARE, and DataMed, in an effort to access more repositories with fewer APIs. In addition, we will create selection criteria for harvesting repositories that are most likely to include MSU-affiliated research datasets. The project team will conduct a survey of researchers at MSU to ask where they publish data, and will include the most commonly-used repositories in our prototype. The Dataset Search homepage will include a form where users can suggest additional repositories to be harvested, or send links to published datasets that were missed by the automatic harvesting.

3.3. Metadata Model

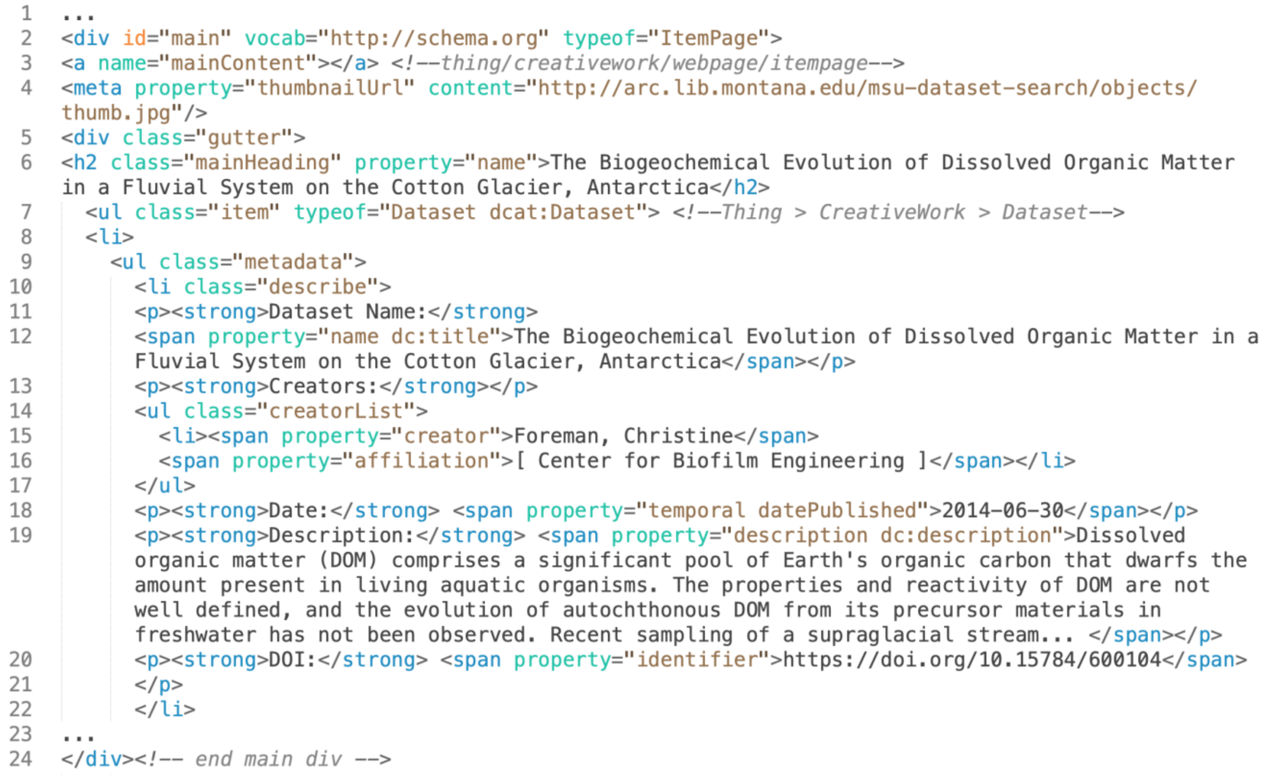
The system includes a metadata model informed by Schema.org [23], DataCite [24], DATS [25], and Project Open Data [26] metadata schemas. For reference, the complete metadata model is available in our MSU Library GitHub repository for the Dataset Search [21]. Our goal was to build the metadata for two purposes: Inventory and discovery. Given the aggregating goals of the Dataset Search, our first metadata goal was to create a data model that allowed us to collect and analyze dataset production by MSU faculty and researchers. A "Creators" table allows us to collect author data and accommodates an arbitrary number of authors for datasets with multiple authors. An "Affiliations" table includes an ID field to help disambiguate our faculty and a discipline/department field to help us understand where common research that involves data is occurring. We also included a "Datasets" table to help catalog and describe the characteristics of the dataset itself. Figure 6, below, shows all of the present fields in the "Datasets" table. Of note here, are the 'dataset_RepositoryName' and 'dataset_doi' fields which allow us to understand the provenance of the dataset and where our faculty and researchers are depositing data. We are also watching for times when we can control the vocabulary of our metadata fields. For example, repository names can be variable, and we are considering a level of human review to standardize this metadata or using the prefix of the DOI [27], which contains a registrant code that could be potentially used to identify the contributing repository and map it to a standardized repository name.



```
CREATE TABLE IF NOT EXISTS `datasets` (
  `recordInfo_recordIdentifier` int(10) NOT NULL COMMENT 'record id',
  `dataset_name` varchar(300) DEFAULT NULL COMMENT 'dataset title',
  `dataset_doi` varchar(300) DEFAULT NULL COMMENT 'original dataset DOI, points at external record',
  `dataset_repositoryName` varchar(255) DEFAULT NULL,
  `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/
  `dataset_spatialCoverage` varchar(30) DEFAULT NULL COMMENT 'geoshape box coordinates OR latitude/long
  `dataset_category1` varchar(255) DEFAULT NULL COMMENT 'linked data category',
  `dataset_category2` varchar(255) DEFAULT NULL COMMENT 'linked data category',
  `dataset_category3` varchar(255) DEFAULT NULL COMMENT 'linked data category',
  `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 Comm
  `dataset_version` varchar(30) DEFAULT NULL COMMENT 'dataset version number',
  `dataset_sameAs` varchar(300) DEFAULT NULL COMMENT 'dataset duplicate content URL for disambiguation
  `dataset_urlHash` varchar(40) DEFAULT NULL COMMENT 'sha1 hash of DOI to help with deduping during har
  `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
  `status` varchar(10) NOT NULL DEFAULT 'r' COMMENT 'record activity status'
) ENGINE=MyISAM DEFAULT CHARSET=utf8;
```

Figure 6. Metadata Fields in "Datasets" table.

Beyond the inventory and analysis goal for the metadata, we looked to create metadata that would help our data become part of commercial search engine indexes. In this instance, we turned to Schema.org, a controlled vocabulary of properties and types created by commercial entities (Google, Bing, Yandex, etc.) that can be embedded in HTML to allow for search engine indexing, to inform our data model. The "Dataset" entity [28] was our guiding principle here and you can see fields like 'dataset_TemporalCoverage' and 'dataset_SpatialCoverage' as a means to qualify the data for search and browse environments. Most importantly, we built a sitemap.xml file [29] to list all of the data items in our catalog and embedded the Schema.org Dataset properties as RDFa structured data [30] in our item pages as seen below in Figure 7.



```
...
<div id="main" vocab="http://schema.org" typeof="ItemPage">
<a name="mainContent"></a> <!--thing/creativework/webpage/itempage-->
<meta property="thumbnailUrl" content="http://arc.lib.montana.edu/msu-dataset-search/objects/thumb.jpg"/>
<div class="gutter">
<h2 class="mainHeading" property="name">The Biogeochemical Evolution of Dissolved Organic Matter in a Fluvial System on the Cotton Glacier, Antarctica</h2>
  <ul class="item" typeof="Dataset dcat:Dataset"> <!--Thing > CreativeWork > Dataset-->
  <li>
    <ul class="metadata">
      <li class="describe">
      <p><strong>Dataset Name:</strong>
      <span property="name dc:title">The Biogeochemical Evolution of Dissolved Organic Matter in a Fluvial System on the Cotton Glacier, Antarctica</span></p>
      <p><strong>Creators:</strong></p>
      <ul class="creatorList">
        <li><span property="creator">Foreman, Christine</span>
        <span property="affiliation">[ Center for Biofilm Engineering ]</span></li>
      </ul>
      <p><strong>Date:</strong> <span property="temporal datePublished">2014-06-30</span></p>
      <p><strong>Description:</strong> <span property="description dc:description">Dissolved organic matter (DOM) comprises a significant pool of Earth's organic carbon that dwarfs the amount present in living aquatic organisms. The properties and reactivity of DOM are not well defined, and the evolution of autochthonous DOM from its precursor materials in freshwater has not been observed. Recent sampling of a supraglacial stream... </span></p>
      <p><strong>DOI:</strong> <span property="identifier">https://doi.org/10.15784/600104</span>
      </p>
      </li>
...
</div><!-- end main div -->
```

Figure 7. The RDFa of data item page.

This sitemap and the RDFa structured data will allow search engines to index our Dataset Search items and we have begun benchmarking the search indexes of Google, Google Dataset Search, and Bing to record the results.

3.4. Unique Metadata Generation

During harvesting, the team also recognized one of the primary limitations of our processing: Limited and variable metadata from the source feeds. There were times during harvest where we noted missing dates, abridged titles, or limited descriptions. In Section 3.1 above, we spoke to a set of seven key metadata fields that we built into our initial metadata records and our interest in supplementing records when data were missing. We also found another unique problem inherent to datasets: They do not exist as textual narratives to analyze. Tabular data tends toward the numerical, especially in the STEM disciplines. With both of these conditions in mind, the team talked through some additional means of creating metadata from datasets and came up with an experimental approach to conduct topic modeling on the dataset creator after the initial harvest and creation of a metadata record. The process we followed is described in detail below.

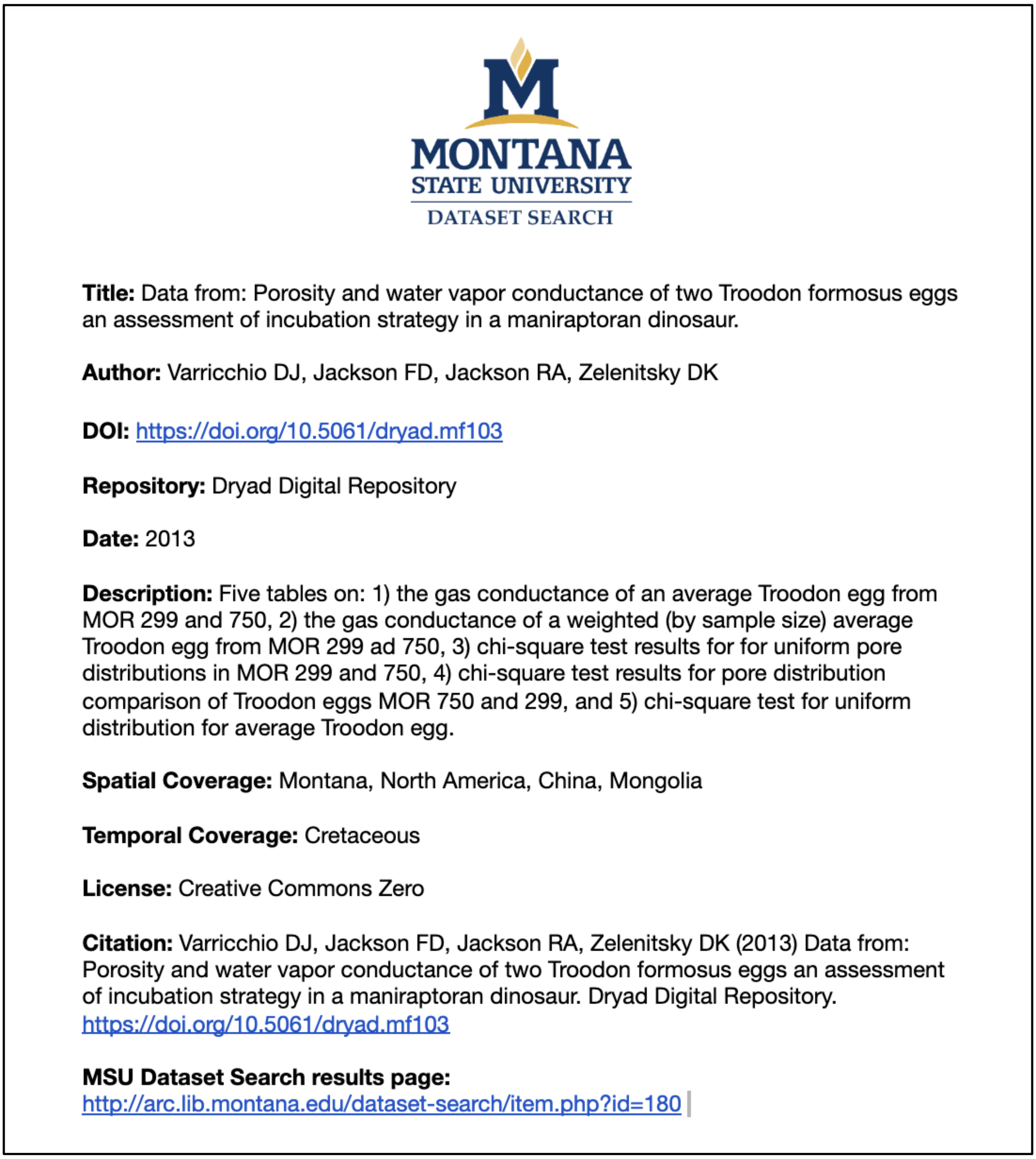
1. Creators could be matched to various academic social networking websites, such as ORCID.org, LinkedIn, or Google Scholar Profiles.
2. Once matched, each of these scholarly profiles is harvested using a web scraping routine and converted into a "bag of words" for topical analysis.
3. The profiles are analyzed and the topics are derived with the top 5 topics becoming an initial set of subject keywords stored in the schema.org "Keywords" property within the metadata record.

As of this writing, this automated metadata generation shows some potential, but we will continue to test and work through the utility of the method. We have some additional ideas about how entity recognition on these same scholarly profiles could be applied to populate our 'dataset_category' fields with linked data. All of these threads are potentially part of a metadata improvement solution. As we refine the methods, we plan on releasing the scripting routine as part of the GitHub repository [21].

3.5. Search Engine Optimization

To promote discovery of Dataset Search metadata for commercial web search engines, we will follow the steps outlined in Arlitsch and O'Brien [31], intended to be used to optimize search engine indexing for institutional repositories. First, sitemaps will be submitted via Google Search Console (Formerly Google Webmaster Tools). Then, through Search Console analysis, errors generated during Google crawls will be identified and improvements will be implemented such as: Improving server performance; implementing unique title and description tags containing the paper's name and abstract; implementing "rel = canonical" tags indicating the preferred URL of each digital object. Second, metadata will be tailored to Google Scholar inclusion guidelines, including: Mapping data repository metadata to Google-supported Highwire Press tags; adding Highwire press meta tags to each index item page.

Google also avoids harvesting sites that it perceives to be "link schemes" [32]. As an index site that provides metadata and links to research data in third-party data repositories, Dataset Search may appear to be a link scheme to Google crawlers. Google suggests two workarounds, intended to help facilitate pay-per-click advertising: Adding a "rel = nofollow" attribute to the <a> tag; redirecting the links to an intermediate page that is blocked from search engines with a robots.txt file. We plan to use a third strategy to promote content to Google crawlers: Autocreating a PDF cover page for each record harvested by the Dataset Search tool. The cover page will contain research dataset metadata and a link to both the third-party data repository and to the Dataset Search results page. Figure 8 shows a mockup of a sample dataset cover page.



MONTANA
STATE UNIVERSITY
DATASET SEARCH

Title: Data from: Porosity and water vapor conductance of two Troodon formosus eggs an assessment of incubation strategy in a maniraptoran dinosaur.

Author: Varricchio DJ, Jackson FD, Jackson RA, Zelenitsky DK

DOI: https://doi.org/10.5061/dryad.mf103

Repository: Dryad Digital Repository

Date: 2013

Description: Five tables on: 1) the gas conductance of an average Troodon egg from MOR 299 and 750, 2) the gas conductance of a weighted (by sample size) average Troodon egg from MOR 299 ad 750, 3) chi-square test results for for uniform pore distributions in MOR 299 and 750, 4) chi-square test results for pore distribution comparison of Troodon eggs MOR 750 and 299, and 5) chi-square test for uniform distribution for average Troodon egg.

Spatial Coverage: Montana, North America, China, Mongolia

Temporal Coverage: Cretaceous

License: Creative Commons Zero

Citation: Varricchio DJ, Jackson FD, Jackson RA, Zelenitsky DK (2013) Data from: Porosity and water vapor conductance of two Troodon formosus eggs an assessment of incubation strategy in a maniraptoran dinosaur. Dryad Digital Repository. https://doi.org/10.5061/dryad.mf103

MSU Dataset Search results page:
http://arc.lib.montana.edu/dataset-search/item.php?id=180

Figure 8. A mockup of a sample PDF dataset cover page.

These PDF cover pages will be stored alongside the Dataset Search metadata, requiring more storage space than the simple metadata files. The team will continue to weigh these risks (increased storage requirements) and benefits (search engine optimization) once the Dataset Search tool goes live.

4. Limitations and Challenges

Over the course of the development of the Dataset Search prototype thus far, we have identified several limitations and challenges.

4.1. API Harvesting

Many data repositories provide an application programming interface (API) for harvesting instead of a feed. To guide the harvesting process from data repositories that employ APIs, we are planning to implement an XML mapping file similar to the one used for feeds. However, unlike feeds, APIs do not have a standardized structure. This could prove to be problematic. There are many possible API methods that could be employed by data repositories, and therefore many different XML maps and workflows that would need to be implemented in order to harvest APIs.

4.2. Institutional Affiliation

Another key challenge for this project is that data repositories may not require disclosure of institutional affiliation. Without institutional affiliation in the data repository metadata, it is difficult to automatically harvest institution-specific content. We are able to obtain a list of researcher names from MSU's Office of Planning and Analysis, and we hope that we can harvest MSU-affiliated datasets by searching for these names, then using the researchers' disciplines to filter the results. ORCID usage is also increasing, and could provide a partial solution; the data repository Zenodo is one notable adopter of ORCID integration.

4.3. Completeness of Content

The two challenges discussed above lead to a third challenge. Ideally, Dataset Search would index every research dataset available from researchers at our institution. However, the challenges with API harvesting and discovering institution-specific datasets will likely prevent the Dataset Search tool from building a fully comprehensive index. However, the Dataset Search will provide a representative sample of datasets from our institution that can be analyzed and inventoried.

4.4. Topic Modeling

We also recognize the limitations behind the methods inherent to our generation of metadata using topic modeling. Our inference of topics from academic social networks we scrape offer a potentially new view into the subjects that make up our datasets, but there are some concerns over how this a priori harvest of researcher profiles could lead to metadata that is less precise and even describing the researcher rather than the dataset. A supplementary quality control method here would be to test this initial metadata topic generation against a topic modeling of the completed metadata records for the datasets themselves. We plan to continue this line of thought and run some topic model testing, but we do maintain that even topics that modify a researcher's primary interests will be of some value in inventory and discovery settings. Even further, researchers have noted how topic modelling methods apply a naïve model of a text as a collection of words decoupled from their syntactic and grammatical contexts of use [33]. We can address this naïve model when we look at the generated topics and consider them in the context of the researcher profile documents (e.g., introduce an element of metadata quality control for these topics before finalization of the descriptive record). All of these challenges are within our research scope and we will look to address these questions whenever it makes sense in the project development. In the end, we still view the topic model method as valid and providing significant enhancement to our metadata generation and automation.

4.5. Local Data Publishing Needs

Another challenge for libraries is that institutional data repositories are often used to archive institutional research data that does not neatly fit into disciplinary data repositories—for example, student research data or very large datasets. Since the Dataset Search tool is not a data repository, but rather a metadata index, it is not designed to store local datasets. This complexity requires multiple solutions, depending on the data itself. For example, student data can be archived in data repositories like Zenodo or Figshare, which are free of cost and have broad collecting policies that can support a wide range of submissions. And the resources saved by foregoing building a local data repository could allow MSU to pay into membership programs for data repositories like Dryad, and to subsidize the cost of archiving large datasets in third-party repositories that can support publishing large datasets—for example, Dryad, Dataverse, and Figshare can publish more than 1TB of data for an additional fee [34].

5. Future Work and Implications

The Dataset Search project is a work-in-progress, and our team is only at the midpoint of our development timeline. The current work of the project is focused on creating the feed and API harvesting workflows for the prototype. In the future, we hope to expand the project to create custom reporting for different MSU departments, develop user access for faculty/students to be able create and update their own records in the system, and to build a JSON-LD API to enable structured data reuse for developers [35]. We also plan to share the prototype more widely with a package installer for local installations.

The project team is also in conversation with university IT about providing metadata records for datasets at our institution that are not publicly available. Providing access to such "invisible" datasets was identified as a challenge by Read et al. in 2015 [36]; the Data Catalog Collaboration Project [37] is one example of a project whose mission includes providing metadata records for nonpublic datasets. In the future, we hope that the Dataset Search tool will be able to provide metadata records as a discovery point for in-progress, sensitive, or otherwise restricted datasets. This feature would provide broader discovery and access to all datasets created by researchers at our institution, even those that are not available in data repositories.

The Dataset Search idea also has implications beyond the local tool that our team is developing. With wider adoption, Dataset Search could lead to three key impacts: (1) Increased discovery and reuse of academic research data; (2) promotion of research data as a scholarly product; and (3) potentially leveraging economies of scale through community-wide implementation. These impacts are discussed in more detail below.

First, an exploratory study conducted by the first author suggests that 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 [38]. The Dataset Search tool will allow research data to be published in subject-specific repositories while additionally being discoverable in a local index. This would promote increased discovery, reuse, and citation of academic research data, ultimately leading to data reuse and potentially increased citations to associated articles [39].

Second, the Dataset Search tool will reinforce the idea that research data is a legitimate scholarly product, both by interoperating with institutional research information management systems, and by creating a public interface where institutions can showcase data as a scholarly product.

Lastly, in a time of tight budgets in universities, economies of scale become increasingly important. The Dataset Search prototype has the potential to be adopted community-wide, creating a single system that can be used across academic libraries while requiring less resource expenditure. 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 university, our profession can make bigger, better systems that promote an important library mission—to provide discovery and access for scholarly products.

Author Contributions: The Dataset Search project is a collaborative project between all four authors. Sara Mannheimer (S.M.) is project lead and project manager, Jason A. Clark (J.A.C.) is metadata lead, James Espeland (J.E.) is software engineer, and Kyle Hagerman (K.H.) is an undergraduate student software development research assistant. Additional contributions, according to the CRediT taxonomy: Conceptualization, S.M., J.A.C., J.E.; software, J.E., K.H.; writing—original draft preparation, S.M., J.A.C., J.E., K.H.; writing—review and editing, S.M., J.A.C.; project administration, S.M.; funding acquisition, S.M.

Funding: 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 www.imls.gov.

Acknowledgments: The authors would like to thank the Dataset Search Advisory Board for their valuable guidance during this project: Kevin Read, Nicole Contaxis, Jeffrey Grethe, and Cynthia Hudson-Vitale.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Newton, M.P.; Miller, C.C.; Bracke, M.S. Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force. *Collect. Manag.* **2010**, *36*, 53–67. doi:10.1080/01462679.2011.530546
2. Johnston, L.; Carlson, J.; Hswe, P.; Hudson-Vitale, C.; Imker, H.; Kozlowski, W.; Olendorf, R.K.; Stewart, C. Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services. *J. eSci. Librariansh.* **2017**, *6*, e1102. doi:10.7191/jeslib.2017.1102
3. Fallaw, C.; Dunham, E.; Wickes, E.; Strong, D.; Stein, A.; Zhang, Q.; Rimkus, K.; Ingram, B.; Imker, H.J. Overly Honest Data Repository Development. *Code4Lib J.* **2016**. http://journal.code4lib.org/articles/11980 (accessed on 12 February 2019).
4. DSpace: The software of choice for academic, non-profit & commercial organizations building open digital repositories. Available online: https://duraspace.org/dspace (accessed on 12 February 2019).
5. Bepress. Digital Commons. Available online: https://www.bepress.com/products/digital-commons (accessed on 12 February 2019).
6. Xie, Z.; Speer, J.; Chen, Y.; Jiang, T.; Brittle, C.; Mather, P. Developing Institutional Research Data Repository: A Case Study. In Proceedings of the Digital Libraries: Knowledge, Information, and Data in an Open Access Society; Morishima, A., Rauber, A., Liew, C.L., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 51–56. doi:10.1007/978-3-319-49304-6_7
7. The Dataverse Project. Open source research data repository software. Available online: https://dataverse.org (accessed on 12 February 2019).
8. CKAN. CKAN Open Source data portal platform. Available online: https://ckan.org (accessed on 12 February 2019).
9. DKAN. DKAN Open Data Platform. Available online: http://www.getdkan.org (accessed on 12 February 2019).
10. Samvera. Samvera: an open source repository solution for digital content. Available online: https://samvera.org (accessed on 12 February 2019).
11. Figshare. Figshare for Institutions. Available online: https://knowledge.figshare.com/institutions (accessed on 12 February 2019).
12. Tind. Tind, a CERN Spinoff, **2019**. Available online: https://tind.io (accessed on 12 February 2019).
13. Mannheimer, S.; Yoon, A.; Greenberg, J.; Feinstein, E.; Scherle, R. A balancing act: The ideal and the realistic in developing Dryad's preservation policy. *First Monday* **2014**, *19*. doi:10.5210/fm.v19i8.5415
14. re3data. Data repositories filtered by country: United States of America. Available online: https://www.re3data.org/search?query=&countries%5B%5D=USA (accessed on 12 February 2019).
15. Mannheimer, S.; Clark, J.A.; Espeland, J. A Prototype for the Institutional Research Data Index. *Zenodo* **2018**. doi:10.5281/zenodo.2561503
16. Lamb, I.; Larson, C. Shining a Light on Scientific Data: Building a Data Catalog to Foster Data Sharing and Reuse. *Code4Lib J.* **2016**, *32*. http://journal.code4lib.org/articles/11421 (accessed on 5 April 2019).
17. Read, K.; Athens, J.; Lamb, I.; Nicholson, J.; Chin, S.; Xu, J.; Rambo, N.; Surkis, A. Promoting Data Reuse and Collaboration at an Academic Medical Center. *Int. J. Digit. Curation* **2015**, 10, 260–267. doi:10.2218/ijdc.v10i1.366
18. Ohno-Machado, L.; Sansone, S.A.; Alter, G.; Fore, I.; Grethe, J.; Xu, H.; Gonzalez-Beltran, A.; Rocca-Serra, P.; Gururaj, A.E.; Bell, E.; et al. Finding useful data across multiple biomedical data repositories using DataMed. *Nat. Genet.* **2017**, *49*, 816–819. doi:10.1038/ng.3864
19. Association of Research Libraries. SHARE notification system project plan. Available online: http://www.arl.org/storage/documents/publications/share-notification-system-project-plan.pdf (accessed on 12 February 2019).
20. Sterman, L.; Clark, J. Citations as Data: Harvesting the Scholarly Record of Your University to Enrich Institutional Knowledge and Support Research. *College Res. Libr.* **2017**, *78*, 952. doi:10.5860/crl.78.7.952
21. Espeland, J.; Clark, J.A.; Hagerman, K.; Mannheimer, S. Code for the IMLS funded MSU Dataset Search. Available online: https://github.com/msulibrary/dataset-search (accessed on 12 February 2019).

22. Montana State University. A Prototype for an Institutional Research Data Index. Funded by the Institute of Museum and library Services LG-89-18-0225-18. Available online: https://www.imls.gov/grants/awarded/lg-89-18-0225-18 (accessed on 22 March 2019).
23. Schema.org. A collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond. Available online: https://schema.org (accessed on 12 February 2019).
24. DataCite. DataCite Metadata Schema. Available online: https://schema.datacite.org (accessed on 12 February 2019).
25. Sansone, S.-A.; Gonzalez-Beltran, A.; Rocca-Serra, P.; Alter, G.; Grethe, J.S.; Xu, H.; Fore, I.M.; Lyle, J.; Gururaj, A.E.; Chen, X.; et al. DATS, the data tag suite to enable discoverability of datasets. *Sci. Data* **2017**, *4*, 170059. doi:10.1038/sdata.2017.59
26. Project Open Data. Metadata Schema v1.1. Available online: https://project-open-data.cio.gov/v1.1/schema (accessed on 12 February 2019).
27. International DOI Foundation. 2 Numbering. In *DOI Handbook*; 2017. Available online: https://www.doi.org/doi_handbook/2_Numbering.html#2.2.2 (accessed on 3 April 2019).
28. Schema.org. Dataset. Available online: http://schema.org/Dataset (accessed on 12 February 2019).
29. Sitemaps.org. What are sitemaps?. https://www.sitemaps.org (accessed on 12 February 2019).
30. W3C. RDFa 1.1 Primer—Third Edition: Rich Structured Data Markup for Web Documents. Available online: https://www.w3.org/TR/rdfa-primer (accessed on 12 February 2019).
31. Arlitsch, K.; O'Brien, P.S. Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Libr. Hi Tech* **2012**, *30*, 60–81. doi:10.1108/07378831211213210
32. Google. Link Schemes. Google Search Console Help. Available online: https://support.google.com/webmasters/answer/66356?hl=en (accessed on 12 February 2019).
33. Brookes, G.; McEnery, T. The Utility of Topic Modelling for Discourse Studies: A Critical Evaluation. *Discourse Studies* **2019**, *21*, 3–21. doi:10.1177/1461445618814032
34. Google. Understand how structured data works. Available online: https://developers.google.com/search/docs/guides/intro-structured-data (accessed on 12 February 2019).
35. Scientific Data. Recommended Data Repositories. Available online: https://www.nature.com/sdata/policies/repositories (accessed on 12 February 2019).
36. Read, K.B.; Sheehan, J.R.; Huerta, M.F.; Knecht, L.S.; Mork, J.G.; Humphreys, B.L.; NIH Big Data Annotator Group Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. *PLOS ONE* **2015**, *10*, e0132735. doi:10.1371/journal.pone.0132735
37. DCCP. Data Catalog Collaboration Project: A Cross-Institutional Collaboration to Index Biomedical Research Data. Available online: https://www.datacatalogcollaborationproject.org (accessed on 22 March 2019).
38. Mannheimer, S.; Sterman, L.; Montana State University-Bozeman; Borda, S. Discovery and Reuse of Open Datasets: An Exploratory Study. *J. eSci. Librariansh.* **2016**, *5*, e1091. doi:10.7191/jeslib.2016.1091
39. Piwowar, H.A.; Vision, T.J. Data reuse and the open data citation advantage. *PeerJ* **2013**, *1*, e175. doi:10.7717/peerj.175