

A **Prototype** for an Institutional Research Data Index

Sara Mannheimer, Jason A. Clark, and
James Espeland

Montana State University

Sara Mannheimer
Data Librarian
Montana State University
[@saramannheimer](#)

Jason A. Clark
Head of Special Collections & Archival Informatics
Montana State University
[@jaclark](#)

James Espeland
Software Engineer
Montana State University
james.espeland@montana.edu

OUTLINE

- 1. The Institutional Data Repository Problem**
- 2. Our Proposed Solution**
- 3. Metadata Model**
- 4. A Very Early Prototype**
- 5. Some Challenges**

1.

THE INSTITUTIONAL DATA REPOSITORY PROBLEM

Out-of-the-box **IR systems** aren't designed for research data

Out-of-the-box **IR systems** aren't designed for research data

1. IRs are designed for articles, which are published once, in a final state.

Out-of-the-box **IR systems** aren't designed for research data

1. IRs are designed for articles, which are published once, in a final state.
2. IRs often don't have data-specific features like file-level description and data-specific metadata.

Out-of-the-box **IR systems** aren't designed for research data

1. IRs are designed for articles, which are published once, in a final state.
2. IRs often don't have data-specific features like file-level description and data-specific metadata.
3. Extensive customizations may prevent upgrading to new versions.

Data-specific repository systems can be resource-intensive

Data-specific repository systems can be resource-intensive

1. Open source systems require developer hours and data storage infrastructure.

Data-specific repository systems are resource-intensive

1. Open source systems require developer hours and data storage infrastructure.
2. Vendor solutions require subscription payments.

Data-specific repository systems can be resource-intensive

1. Open source systems require developer hours and data storage infrastructure.
2. Vendor solutions require subscription payments.
3. Data repository resources are expended in addition to those required for existing IRs.

There are currently
~1000 data repositories
in the United States
&
>2000 data repositories
worldwide

*source: [re3data](#)

2.

OUR PROPOSED SOLUTION

7-19-2016

Discovery and Reuse of Open Datasets: An Exploratory Study

Sara Mannheimer

Montana State University-Bozeman

Leila Belle Sterman

Montana State University-Bozeman

Susan Borda

Montana State University-Bozeman

<https://doi.org/10.7191/jeslib.2016.1091>

CONSERVE RESOURCES + INCREASE DISCOVERY

Our results suggest that research data are more likely to be discovered and reused if they are:

1. Archived in a discipline-specific repository
2. Indexed in multiple places online

CONSERVE RESOURCES + INCREASE DISCOVERY

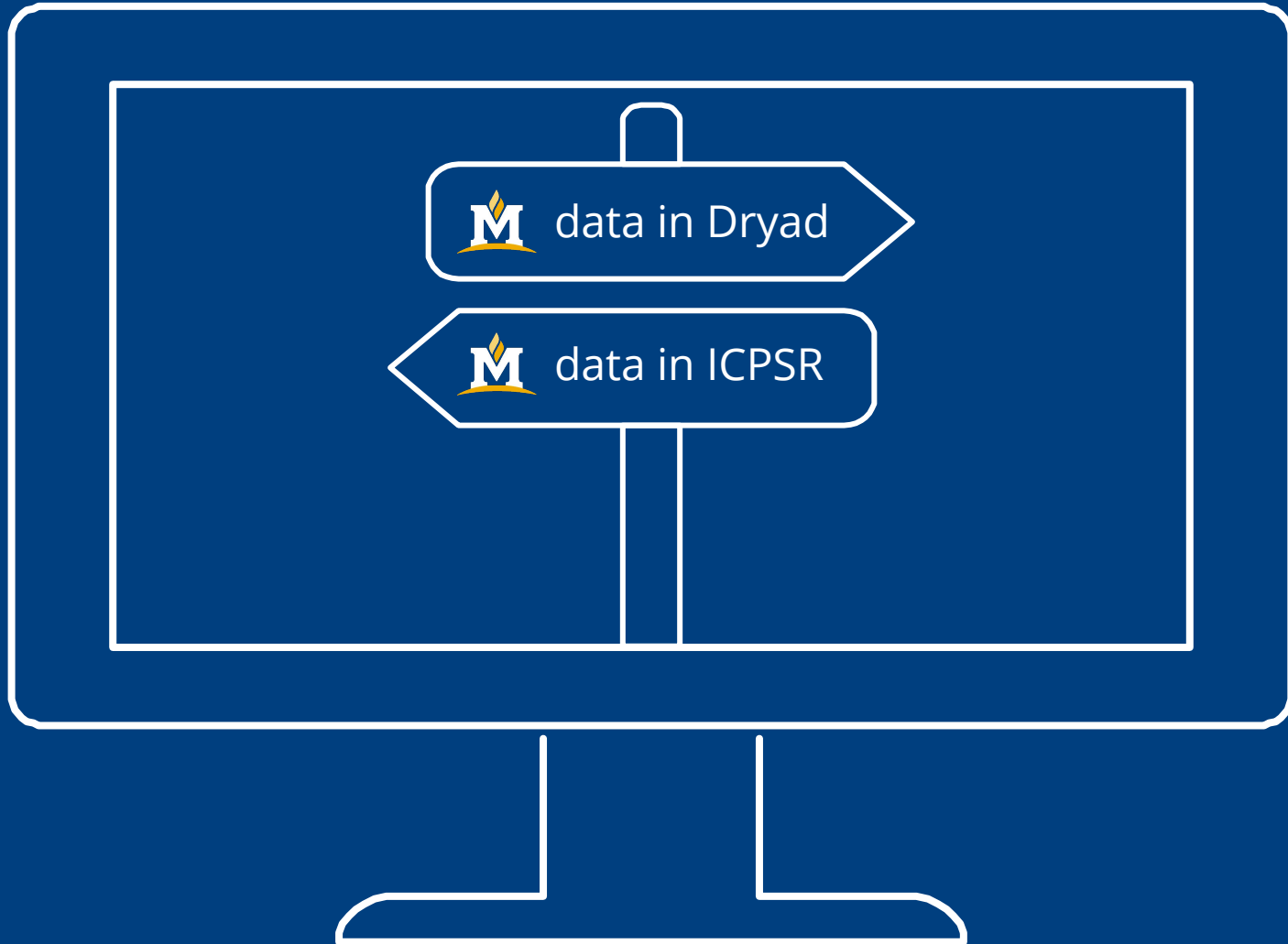
1. Direct researchers to deposit research data in third-party repositories
2. Provide local access to institutional research datasets through a research data index

INSTITUTIONAL **RESEARCH DATA** INDEX

IRDI is a metadata index that promotes discovery and reuse of institutional research datasets while expending far fewer resources than those required for an institutional data repository.



INSTITUTIONAL **RESEARCH DATA** INDEX



IRDI **INSPIRATIONS**

DataMed

NYU Health Sciences Library Data Catalog

Cinergi

OmicsDI

SHARE

IRDI INNOVATIONS

1. Institutional focus
2. Automatic generation of new descriptive metadata for individual datasets
3. SEO

3.

METADATA MODEL

METADATA CHOICES:

1. Schema.org Datasets Type
2. DCAT
3. BioCaddie
4. CSV on the Web Working Group

METADATA for Discovery:



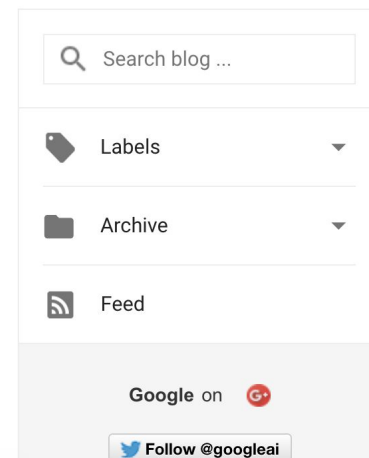
The latest news from Google AI

Facilitating the discovery of public datasets

Tuesday, January 24, 2017

Posted by Natasha Noy, Google Research and Dan Brickley, Open Source Programs Office

There are many hundreds of data repositories on the Web, providing access to tens of thousands—or millions—of datasets. National and regional governments, scientific publishers and consortia, commercial data providers, and others [publish data](#) for fields ranging from [social science](#) to [life science](#) to [high-energy physics](#) to [climate science](#) and more. Access to this data is critical to facilitating reproducibility of research results, enabling scientists to build on others' work, and providing data journalists easier access to information and its provenance. For these reasons, many publishers and funding agencies now require that scientists make their research data available publicly.



<https://ai.googleblog.com/2017/01/facilitating-discovery-of-public.html>

METADATA for Discovery:

Google Structured Data Testing Tool



1



NEW TEST



```
1 <!doctype html>
2 <html lang="en" vocab="http://schema.org/" typeof="WebPage">
3 <head>
4 <title>MSU Data Catalog Record - Montana State University (MSU)
  Library</title>
5 <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
6 <meta name="description" content="Dataset metadata" />
7 <meta name="keywords" content="" />
8 <meta name="robots" content="noindex, nofollow" />
9 <link rel="canonical" href="http://shelf.lib.montana.edu/msu-data-
  catalog/item/1" />
10 <link rel="stylesheet" type="text/css" media="screen, projection"
  href="/msu-data-catalog/meta/styles/master.css" />
11 </head>
12 <body class="fullWidth default">
13 <div id="contain">
14 <div id="mastHead">
15   <h1><a href="/msu-data-catalog/index.php">Montana State University
    Data Catalog - Montana State University (MSU) Library</a></h1>
16   <p class="hidden"><a href="#mainContent" title="main content">skip
    navigation</a></p>
17 </div><!-- end masthead div -->
18 <div id="main" vocab="http://schema.org" typeof="ItemPage">
19   <a name="mainContent"></a> <!--
  thing/creativeWork/webpage/itempage-->
20   <meta property="thumbnailUrl"
  content="http://shelf.lib.montana.edu/msu-data-
  catalog/objects/thumb.jpg" />
21   <div class="gutter">
```

← ItemPage

All (1) ▾

ItemPage

0 ERRORS 0 WARNINGS ^

@type

ItemPage

thumbnailUrl

http://shelf.lib.montana.edu/msu-data-
catalog/objects/thumb.jpg

name

Data from: Entomopathogenic Nematodes
Combined with Adjuvants Presents a New
Potential Biological Control Method for
Managing the Wheat Stem Sawfly, *Cephus
cinctus* (Hymenoptera: Cephidae)

METADATA TABLES:

4 tables

 Structure	 SQL	 Search	 Query	
Table 	Action			
☐ affiliations	 Browse	 Structure	 Search	 Insert
☐ creators	 Browse	 Structure	 Search	 Insert
☐ datasets	 Browse	 Structure	 Search	 Insert
☐ feeds	 Browse	 Structure	 Search	 Insert
4 tables	Sum			

What is a **DATASET**?

1. A single table?
2. A collection of related tables?
3. A protein sequence?
4. A set of images?
5. An API that provides access to data?

What is a DATASET?

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

METADATA **FIELDS**:

1. Set up for deduplication (sha256 hash)
2. Set up for resolution (DOIs)
3. Set up for reuse (APIs)

METADATA IDENTITIES:

- Disambiguating Researchers

John Clark at PSU

Vs

John Clark at MSU

METADATA IDENTITIES:

Can we find a “Topical Fingerprint” for each researcher to automate topical metadata?

METADATA IDENTITIES:

 [jasonclark](#) / [linked-people-bioscience](#)

 Unwatch ▾ 2  Star 5  Fork 2

 Code

 Issues 7

 Pull requests 0

 Projects 0

 Wiki

 Insights

 Settings

Branch: master ▾

[linked-people-bioscience](#) / [meta](#) / [scripts](#) /

Create new file

Upload files

Find file

History

 **jasonclark** loaded basic app shell and components for #mozspint

Latest commit c2665c9 on May 31, 2017

..

 getGoogleScholarCitationsSkills.py	set default URL to harvest	a year ago
 getLinkedInSkills.py	set default URL to harvest	a year ago
 getMSAcademicSearchKeywords.py	loaded basic app shell and components for #mozspint	a year ago
 getOrcidKeywords.py	Initial commit	a year ago

METADATA IDENTITIES:

```
linked-people-profile.html • <iframe width="560" height="315" src="https://www. bot.js x getOrcidKeywords.py • list.php • index.html x #HowToJason x
2
3 from bs4 import BeautifulSoup
4 import csv
5 import json
6 import os
7 import requests
8 import sys
9
10 #URI of the data source
11 URI = sys.argv[1] if len(sys.argv) > 1 else 'http://orcid.org/0000-0002-3588-6257'
12
13 def parse_source(uri):
14     request = requests.get(uri, headers={'User-Agent' : 'Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.
15
16     #check for HTTP codes other than 200
17     if request.status_code != 200:
18         print('Status:', request.status_code, 'Problem with the request. Exiting.')
19         exit()
20
21     soup = BeautifulSoup(request.text, 'html.parser')
22
23     pageTitle = soup.title.string
24     pageFileName = pageTitle.replace(' ', '-').lower()
25
26     print ('Page Title: \n' + pageTitle)
27     #print ('Person: \n' + personName)
28     print ('Page URL: \n' + uri)
```

<https://github.com/jasonclark/linked-people-bioscience/blob/master/meta/scripts/getOrcidKeywords.py>

4.

A VERY EARLY PROTOTYPE

Demo

5. SOME CHALLENGES

- 1. Metadata sources**
- 2. Institutional affiliation**
- 3. Outlier datasets**
- 4. Search engine indexing**

1. METADATA SOURCES

METADATA SOURCES

The challenge

RSS feeds and APIs are unstandardized



METADATA SOURCES

A potential solution

Dataset metadata services such as DataMed, Cinergi, OmicsDI, and DataCite



2. INSTITUTIONAL AFFILIATION

INSTITUTIONAL AFFILIATION

The challenge

Data repositories often don't require institutional affiliation.

Records in IRDI may not show a complete picture of all of the institutional research datasets that have been published.



INSTITUTIONAL AFFILIATION

Some potential solutions

Querying data repositories for the full names of MSU researchers.

This is labor-intensive and fails to account for name disambiguation



INSTITUTIONAL AFFILIATION

Some potential solutions

ORCID integration

But most data repositories still don't use
ORCID



INSTITUTIONAL AFFILIATION

Some potential solutions

Grant numbers—most data repositories provide a metadata field for grant information.

These numbers can be cross-referenced with institutional award information



INSTITUTIONAL AFFILIATION

Some potential solutions

Cross-referencing research datasets with underlying publications (if applicable).

Publications have institutional affiliation information.



3. OUTLIER DATASETS

OUTLIER DATASETS

The challenge

Institutional data repositories may be used to archive institutional research data that doesn't neatly fit into disciplinary data repositories. Since IRDI is a metadata index—not a data repository—it is not designed to store local datasets.



OUTLIER DATASETS

A potential solution

The resources saved by foregoing building a local data repository could allow institutions to subsidize any potential cost to archive unique research data in third-party repositories.



4. SEARCH ENGINE INDEXING

SEARCH ENGINE INDEXING

The challenge

Search engines like Google may not index metadata-only repositories



SEARCH ENGINE INDEXING

A potential solution

Provide an object in IRDI in order to improve search engine optimization.

For example, a downloadable text document with dataset metadata and a link to the external data repository.



FINAL THOUGHTS

Prototype is in early stages

There are challenges ahead

**We'd love feedback from the
Open Repositories community**

THANKS!

+ questions and feedback

@saramannheimer

sara.mannheimer@montana.edu

@jaclark

jaclark@montana.edu

james.espeland@montana.edu

ACKNOWLEDGEMENTS

The IRDI prototype is based on a system developed at MSU Library that uses Javascript and PHP scripts to parse RSS feeds and harvest metadata for institutional publications.

Code: <https://github.com/msulibrary/msu-research-citations>

Article: Sterman & Clark, 2017, <https://doi.org/10.5860/crl.78.7.952>

The design of this presentation is based on the Mercurio template from [Slides Carnival](#)