NM EPSCoR Data Management and Project Data Repositories

Karl Benedict Jonathan Wheeler

February 4, 2020

A High-Level Integration of the Research Lifecycle and a Project Data Preservation and Publication Workflow

The EPSCoR SMART Grid Center includes diverse research activities that are being performed on a variety of systems ranging from desktop workstations and laptop computers to the high-performance computer systems at NMSU and UNM. This work is progressing through a standard research lifecycle that includes the production, management, and analysis of data; the sharing of data with project collaborators to enable shared research activities; and the publication of results. It is in the context of the publication of results that the project's data preservation and publication platforms come into play - providing the capability to preserve the data products generated by the project, and enabling FAIR (Findable, Accessible, Interoperable, and Reusable) access to project data products, including the generation of persistent identifiers (DOIs) for the published data. Meeting this objective meets the requirements that are imposed by both our project sponsor [NSF] and a growing number of publishers who require public access to data associated with published papers. Meeting this need requires the generation of reusable data products, the collection and compilation of documentation (metadata) associated with those data, and the placement of those data in an appropriate repository where they can be discovered and accessed.

In support of the EPSCoR SMART Grid Center project we have a number of current and planned capabilities to enable the efficient management, preservation, and sharing of FAIR data. The following figure highlights the workflow that we currently support and future workflow components that will be developed over the course of the project.

Publishing Data in Dryad

The Dryad data repository is maintained by the California Digital Library. The University of New Mexico is a member institution through the NM ESPCoR SMART Grid Center. There are two separate paths to publication for SMART Grid researchers:

1. **Everyone** may use mediated upload services, supported by the UNM Libraries' Research Data Services. This service is available to both UNM and non-UNM affiliated researchers. Note that currently at least one author must be from UNM, and that person will have to initiate the dataset upload process in Dryad.
2. **UNM affiliated researchers** can directly access Dryad by logging with their ORCiD and UNM credentials.

Regardless of institutional affiliation, the first author listed on a dataset **must have an ORCiD**. This is a Dryad requirement, not an EPSCoR requirement.

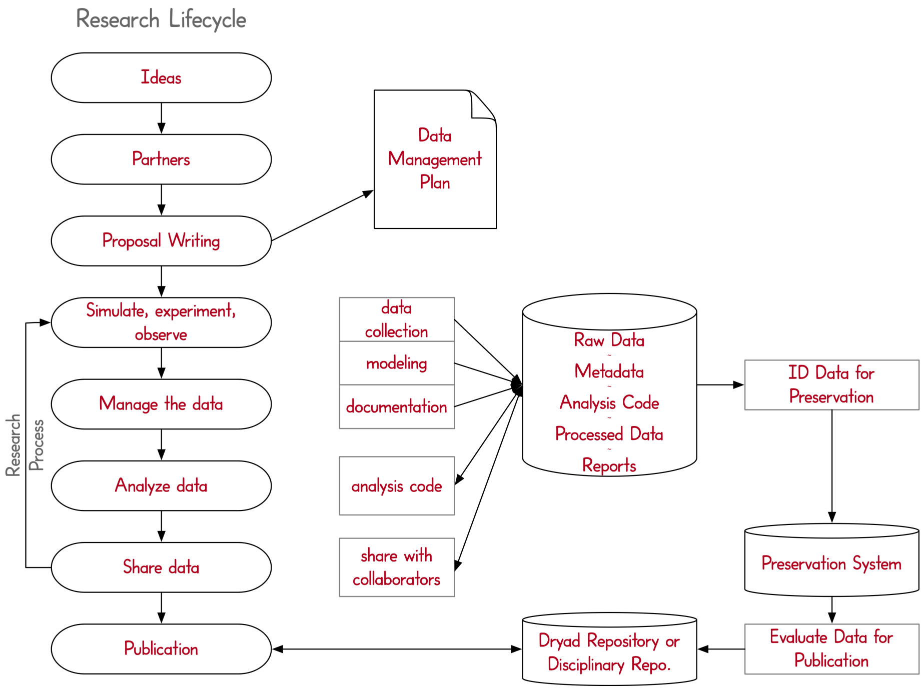


Figure 1: Research lifecycle with high-level data workflow components

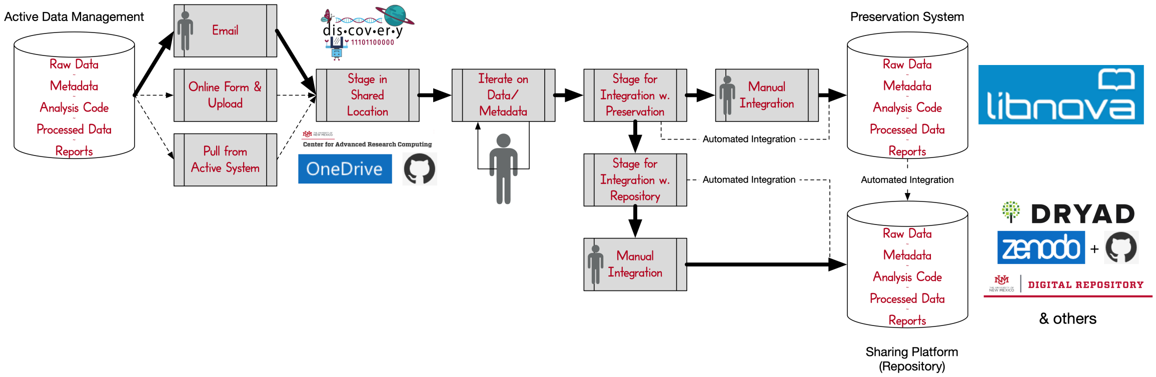
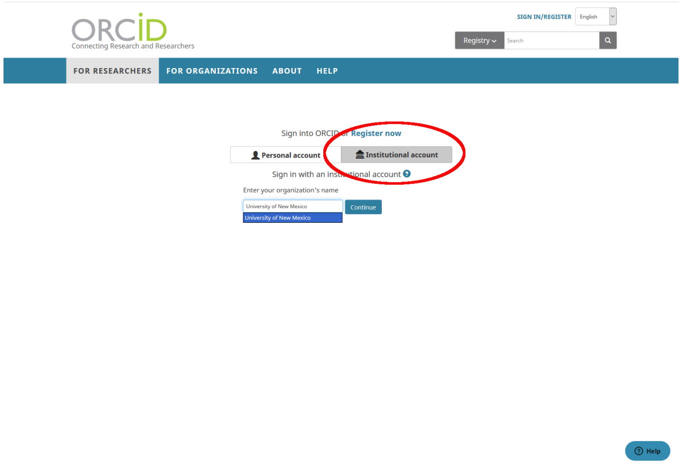


Figure 2: EPSCoR SMART Grid Center data preservation and publication workflow

Getting an ORCiD

ORCiD is a unique, persistent identifier for researchers. Publishers increasingly require ORCiDs as part of the manuscript submission process, but anyone who needs to register with ORCiD can

1. Go to https://orcid.org/. Click on *SIGN IN/REGISTER* at the top right of the page.
2. UNM and NMSU researchers can click on *Institutional account*. Researchers affiliated with other institutions can click on *Personal account*.
3. Researchers can fill out their profile as needed, but for Dryad purposes the ORCiD iD is generated by the system right away.



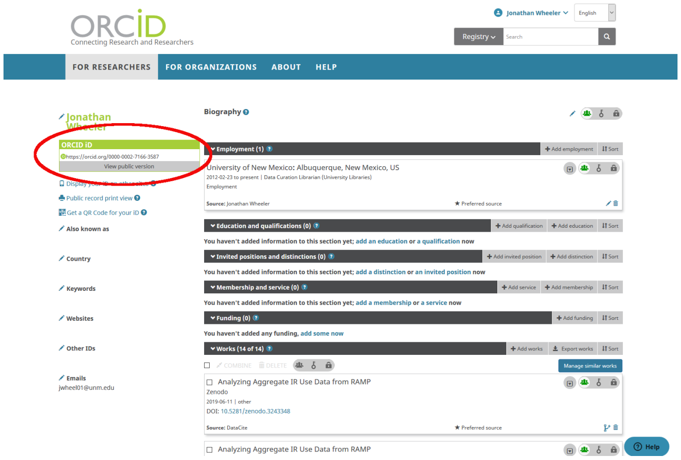
Sign into ORCID or Register now

Sign in with an institutional account

Enter your organization's name: University of New Mexico

University of New Mexico

Figure 3: ORCID sign in



Jonathan Wheeler

ORCID iD

https://orcid.org/0000-0002-7166-3587

Emails

jwheel01@unm.edu

Biography

Employment (1)

University of New Mexico: Albuquerque, New Mexico, US

2012-02-23 to present | Data Curation Librarian (University Libraries)

Employment

Source: Jonathan Wheeler ★ Preferred source

Education and qualifications (0)

You haven't added information to this section yet; add an education or a qualification now

Invited positions and distinctions (0)

You haven't added information to this section yet; add a distinction or an invited position now

Membership and service (0)

You haven't added information to this section yet; add a membership or a service now

Funding (0)

You haven't added any funding, add some now

Works (14 of 14)

☐ Analyzing Aggregate IR Use Data from RAMP

Zenodo

2019-06-11 | other

DOI: 10.5281/zenodo.3243348

Source: DataCite ★ Preferred source

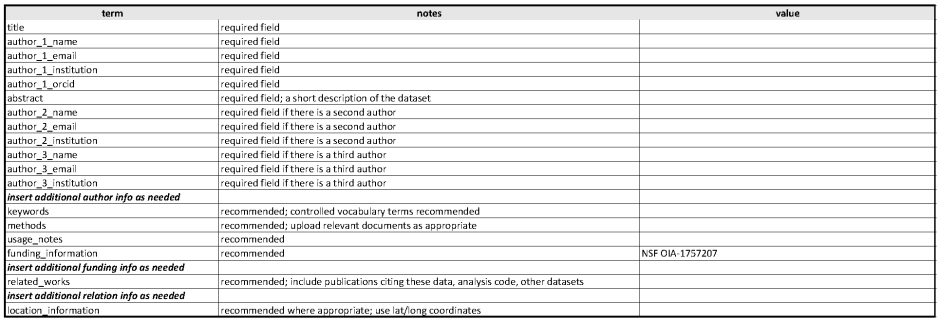
☐ Analyzing Aggregate IR Use Data from RAMP

Figure 4: ORCID iD

Submitting Data to Research Data Services

As noted above, the UNM Libraries' Research Data Services (RDS) unit can mediate all deposits to Dryad on behalf of researchers. An SMART Grid specific online interface for submitting data to RDS is in development. Until that is ready researchers should:

1. Use the provided CSV metadata template to prepare metadata and documentation. The fields in the template correspond to required fields in Dryad and align with the DataCite metadata schema used by Dryad and other data repositories.
2. Contact RDS at rds@unm.edu to request access to a data upload directory. At this time, the default upload directory will be a shared OneDrive directory but RDS will work with researcher preferences. Alternative methods for submitting data are available.
3. The data upload directory will contain a copy of the CSV metadata template. Researchers can edit this or upload a local copy.
4. One UNM affiliated author will log into Dryad to begin the submission process. RDS will coordinate with that author to complete the submission.
5. RDS will consult with researchers as needed for additional metadata or documentaiton and will submit the dataset to Dryad on the researcher's behalf. Datasets will be published under a "review" status to allow for feedback, updates, and corrections from researchers prior to final publication.



term	notes	value
title	required field	
author_1_name	required field	
author_1_email	required field	
author_1_institution	required field	
author_1_orcid	required field	
abstract	required field; a short description of the dataset	
author_2_name	required field if there is a second author	
author_2_email	required field if there is a second author	
author_2_institution	required field if there is a second author	
author_3_name	required field if there is a third author	
author_3_email	required field if there is a third author	
author_3_institution	required field if there is a third author	
insert additional author info as needed		
keywords	recommended; controlled vocabulary terms recommended	
methods	recommended; upload relevant documents as appropriate	
usage_notes	recommended	
funding_information	recommended	NSF OIA-1757207
insert additional funding info as needed		
related_works	recommended; include publications citing these data, analysis code, other datasets	
insert additional relation info as needed		
location_information	recommended where appropriate; use lat/long coordinates	

Figure 5: metadata template

Overview of the Dryad Submission Process

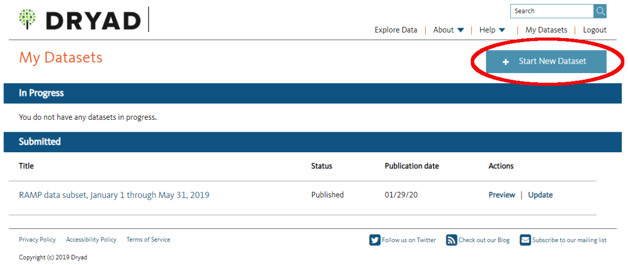
Dataset Size Limits

Dryad provides two options for uploading data. Both options have different size limits:

- When uploading directly from a computer the limit is 10GB per DOI.
- When uploading from a server or shared files service (Google Drive, Box, etc.) the limit is 300GB per DOI.

Note that the limits are per DOI. That means there is a practical limit on the number and size of revisions that may be submitted for any single dataset.

1. After logging in, click the button to *Start a New Dataset*.
2. Indicate whether the data are related to a manuscript, a published article, or "other." The section headed *Dataset: Basic Information* includes the required metadata fields. Note that the submitting author's ORCiD is included. Other authors will receive an email requesting them to add their ORCiD.
3. Additional (recommended) metadata can be provided. Please note that some usage information has been specified project-wide in the DMP.
4. Upload files using one of the two methods described above.
5. The final *Review and Submit* step includes options for enabling private peer review. Note that the default license in Dryad is public domain.



My Datasets

In Progress

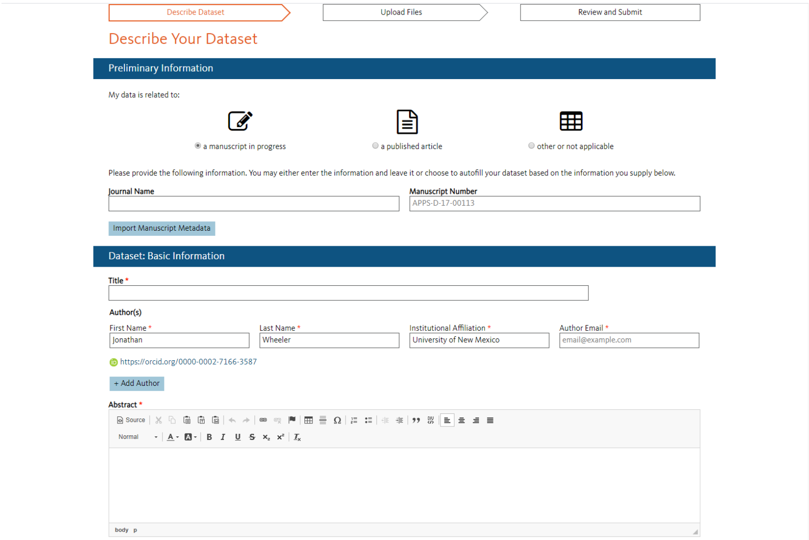
You do not have any datasets in progress.

Submitted

Title	Status	Publication date	Actions
RAMP data subset, January 1 through May 31, 2019	Published	01/29/20	Preview | Update

Copyright (c) 2019 Dryad

Figure 6: Dryad step 1



Describe Your Dataset

Preliminary Information

My data is related to:

● a manuscript in progress ○ a published article ○ other or not applicable

Please provide the following information. You may either enter the information and leave it or choose to autofill your dataset based on the information you supply below.

Journal Name

Manuscript Number APPS-D-17-00113

Dataset: Basic Information

Title *

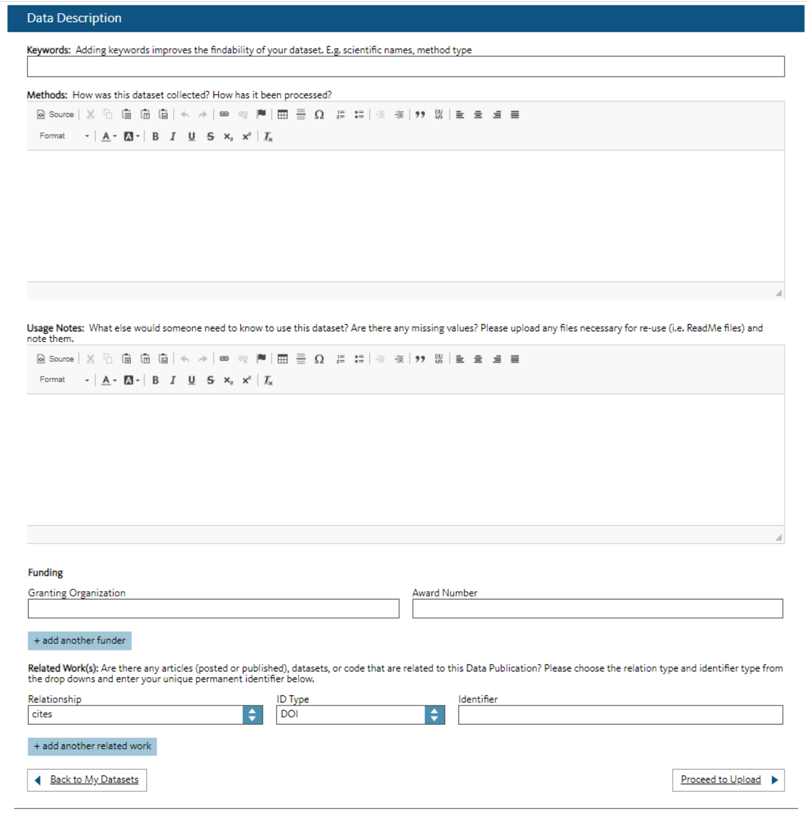
Author(s)

First Name *	Last Name *	Institutional Affiliation *	Author Email *
Jonathan	Wheeler	University of New Mexico	email@example.com

https://orcid.org/0000-0002-7166-3587

Abstract *

Figure 7: Dryad step 2



Data Description

Keywords: Adding keywords improves the findability of your dataset. E.g. scientific names, method type

Methods: How was this dataset collected? How has it been processed?

Usage Notes: What else would someone need to know to use this dataset? Are there any missing values? Please upload any files necessary for re-use (i.e. ReadMe files) and note them.

Funding

Granting Organization

Award Number

Related Work(s): Are there any articles (posted or published), datasets, or code that are related to this Data Publication? Please choose the relation type and identifier type from the drop downs and enter your unique permanent identifier below.

Relationship: cites

ID Type: DOI

Identifier

Figure 8: Dryad step 3

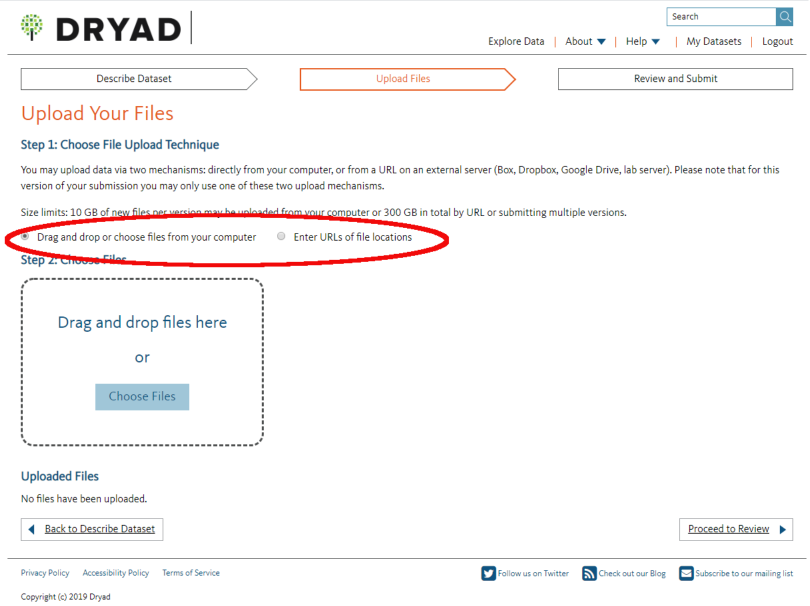
About DOIs

Datasets published in Dryad are assigned a DOI immediately upon submission. No additional workflow is required for a DOI in Dryad.

DOIs are available for datasets and research products that may be published in repositories other then Dryad. Many repositories will create DOIs, but if not UNM is a DataCite member and RDS can manually generate DOIs as needed.

Dataset Versioning

Dryad supports dataset versioning. Any dataset can be revised using a process similar to the one outlined above for submitting a new dataset. Dataset revisions in Dryad all have the same DOI, with changes to files included in the *Data Files* section.



Upload Your Files

Step 1: Choose File Upload Technique

You may upload data via two mechanisms: directly from your computer, or from a URL on an external server (Box, Dropbox, Google Drive, lab server). Please note that for this version of your submission you may only use one of these two upload mechanisms.

Size limits: 10 GB of new files per version may be uploaded from your computer or 300 GB in total by URL or submitting multiple versions.

● Drag and drop or choose files from your computer ○ Enter URLs of file locations

Step 2: Choose Files

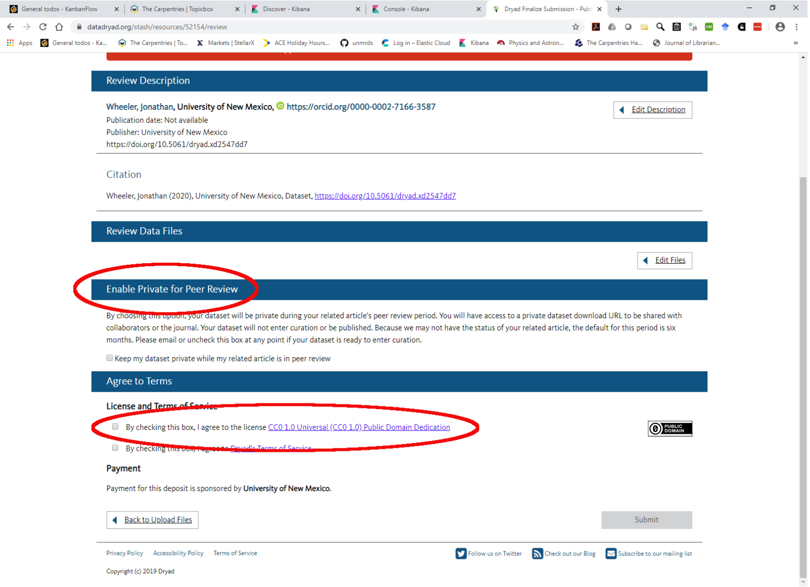
Uploaded Files

No files have been uploaded.

Back to Describe Dataset
Proceed to Review
Privacy Policy Accessibility Policy Terms of Service
Follow us on Twitter Check out our Blog Subscribe to our mailing list

Copyright (c) 2019 Dryad

Figure 9: Dryad step 4


General todos - KanbanFlow
The Carpentries | Topicbox
Discover - Kibana
Console - Kibana
Dryad Finalize Submission - Publ
datadryad.org/stash/resources/52154/review
Apps General todos - Ka... The Carpentries | To... Markets | StellarX ACE Holiday Hours... unmrds Log in – Elastic Cloud Kibana Physics and Astron... The Carpentries Ha... Journal of Librarian...

Review Description

Wheeler, Jonathan, University of New Mexico, https://orcid.org/0000-0002-7166-3587

Publication date: Not available

Publisher: University of New Mexico

https://doi.org/10.5061/dryad.xd2547dd7

Citation

Wheeler, Jonathan (2020), University of New Mexico, Dataset, https://doi.org/10.5061/dryad.xd2547dd7

Review Data Files

Enable Private for Peer Review

By choosing this option, your dataset will be private during your related article's peer review period. You will have access to a private dataset download URL to be shared with collaborators or the journal. Your dataset will not enter curation or be published. Because we may not have the status of your related article, the default for this period is six months. Please email or uncheck this box at any point if your dataset is ready to enter curation.

□ Keep my dataset private while my related article is in peer review

Agree to Terms

License and Terms of Service

□ By checking this box, I agree to the license CC0 1.0 Universal (CC0 1.0) Public Domain Dedication

□ By checking this box, I agree to Dryad's Terms of Service

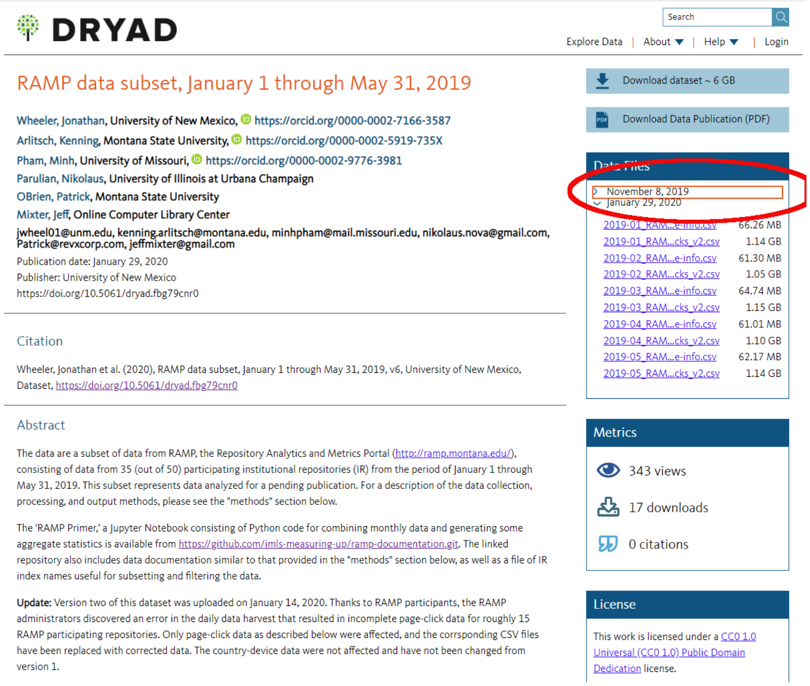
Payment

Payment for this deposit is sponsored by **University of New Mexico.**

PUBLIC DOMAIN
Back to Upload Files
Submit
Privacy Policy Accessibility Policy Terms of Service
Follow us on Twitter Check out our Blog Subscribe to our mailing list

Copyright (c) 2019 Dryad

Figure 10: Dryad step 5


DRYAD
Search
Explore Data | About ▼ | Help ▼ | Login
Download dataset ~ 6 GB
Download Data Publication (PDF)

RAMP data subset, January 1 through May 31, 2019

Wheeler, Jonathan, University of New Mexico, https://orcid.org/0000-0002-7166-3587
Arlitsch, Kenning, Montana State University, https://orcid.org/0000-0002-5919-735X
Pham, Minh, University of Missouri, https://orcid.org/0000-0002-9776-3981
Parulian, Nikolaus, University of Illinois at Urbana Champaign
OBrien, Patrick, Montana State University
Mixter, Jeff, Online Computer Library Center
jwheel01@unm.edu, kenning.arlitsch@montana.edu, minhpham@mail.missouri.edu, nikolaus.nova@gmail.com, Patrick@revxcorp.com, jeffmixter@gmail.com
Publication date: January 29, 2020
Publisher: University of New Mexico
https://doi.org/10.5061/dryad.fbg79cnr0

Citation

Wheeler, Jonathan et al. (2020), RAMP data subset, January 1 through May 31, 2019, v6, University of New Mexico, Dataset, https://doi.org/10.5061/dryad.fbg79cnr0

Abstract

The data are a subset of data from RAMP, the Repository Analytics and Metrics Portal (http://ramp.montana.edu/), consisting of data from 35 (out of 50) participating institutional repositories (IR) from the period of January 1 through May 31, 2019. This subset represents data analyzed for a pending publication. For a description of the data collection, processing, and output methods, please see the "methods" section below.

The 'RAMP Primer,' a Jupyter Notebook consisting of Python code for combining monthly data and generating some aggregate statistics is available from https://github.com/imls-measuring-up/ramp-documentation.git. The linked repository also includes data documentation similar to that provided in the "methods" section below, as well as a file of IR index names useful for subsetting and filtering the data.

Update: Version two of this dataset was uploaded on January 14, 2020. Thanks to RAMP participants, the RAMP administrators discovered an error in the daily data harvest that resulted in incomplete page-click data for roughly 15 RAMP participating repositories. Only page-click data as described below were affected, and the corrsponding CSV files have been replaced with corrected data. The country-device data were not affected and have not been changed from version 1.

Data Files

› November 8, 2019
⌄ January 29, 2020

2019-01_RAM...e-info.csv	66.26 MB
2019-01_RAM...cks_v2.csv	1.14 GB
2019-02_RAM...e-info.csv	61.30 MB
2019-02_RAM...cks_v2.csv	1.05 GB
2019-03_RAM...e-info.csv	64.74 MB
2019-03_RAM...cks_v2.csv	1.15 GB
2019-04_RAM...e-info.csv	61.01 MB
2019-04_RAM...cks_v2.csv	1.10 GB
2019-05_RAM...e-info.csv	62.17 MB
2019-05_RAM...cks_v2.csv	1.14 GB

Metrics

343 views
17 downloads
0 citations

License

This work is licensed under a CC0 1.0 Universal (CC0 1.0) Public Domain Dedication license.

Figure 11: Dryad step 6

Other features to note in the image above include the link to download the entire current version of the dataset, as well as the *Download Data Publication* button, which exports the metadata in PDF format. The PDF includes links to the data files.