

Research Data Camp: Data Publishing and Repositories



Melissa Cantrell, Assistant Professor, Scholarly Communication Librarian

August 20th, 2024



Center for Research Data & Digital Scholarship

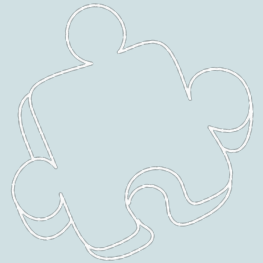
UNIVERSITY OF COLORADO **BOULDER**

Agenda

1. What is data publishing?
2. Why publish data?
3. How to publish data
 - a. Intro to FAIR principles
 - b. Top considerations
 - c. CU Scholar/Dryad examples
4. Questions and wrap-up



1. What is data publishing?



Working definition

- Making research data and metadata/documentation publicly available (or with appropriate access controls) via a formal web-based repository/database
- Preferably in adherence with [FAIR data principles](#) and/or other standards for data, metadata, and repository quality

Related terms

- Data sharing
- Data curation
- Data archiving
- Data preservation

2. Why publish data?



Why publish data?

1. Scientific and public good

- a. Advance scientific innovation
- b. Address reproducibility

2. Journal/publisher requirements

- a. “Data availability statements”
- b. FAIR repositories with citations via persistent IDs

3. Funder requirements

- a. Part of NSF data management plans since 2011
- b. Part of NIH data management and sharing policy since 2023

Thinking Ahead

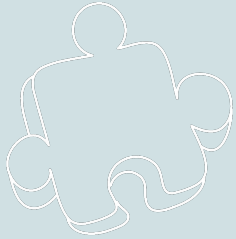


**Open, accessible, and
reproducible data
advancing the public good**

**Data management
planning and
protocols**

**FAIR data
publication
principles**

3. How to publish data



Introduction to FAIR data principles ([Wilkinson et al., 2016](#))

FAIR DATA PRINCIPLES

AH!



FINDABLE



ACCESSIBLE

HOW DO YOU
OPEN A .XZQ FILE?



INTEROPERABLE



REUSABLE

Findable (F)

- Apply a globally unique and persistent identifier
- Describe your data in a data repository

FINDABLE

Unique identifiers and metadata are used to allow data to be located quickly and efficiently



Accessible (A)

- Consider what will be shared, and share via a open, free, and universally implementable protocol
- Metadata are valuable and accessible, even when the data are no longer available

ACCESSIBLE

Data is open, free
and universally
available for
research
discovery efforts



Interoperable (I)

- Use:
 - Open formats
 - Consistent vocabulary
 - Common metadata standards

INTER- OPERABLE

A common programming language is used to allow use in a broad range of applications



Reusable (R)

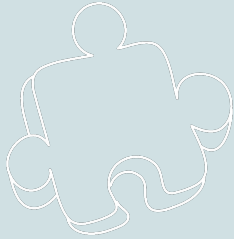
- Origin, context, history, and who to credit/cite are all crucial for data reusability
- Consider permitted use and apply the appropriate license

REUSABLE

All data is clearly described and outlines associated data-use standards



Top Considerations...**Before you start collecting data**



Top Considerations for You # 1

Have clear documentation and a data management plan from square one.

- Think ahead about repositories and requirements for a finished project
- Document throughout the process/project: how data was created/gathered/used/etc.
- R (Reusable) in FAIR is very hard to achieve just at the end of the project; important to think about from the beginning

Top Considerations for You # 1

Before you start collecting data, think about:

- How much of your data will you/can you share?
- How and where will you share your data?
- When will you share your data?
- With whom will you share your data?



Top Considerations for You # 2:

Select a FAIR-aligned data repository

- [CU Scholar](#)
 - FAIR-aligned public access repository for CU Boulder affiliated researchers (i.e., have an IdentiKey)
 - Has [CoreTrustSeal](#) certification
 - Review and curation of all data sets
 - DataCite DOIs registered for all data sets
 - Public access to large data sets via Globus and PetaLibrary
 - Free to deposit up to 500 GB per data set for CU Boulder affiliated researchers
 - Over 1600 data sets published in CU Scholar to date



DRYAD

Top Considerations for You # 2:

Select a FAIR-aligned data repository

- [Dryad @ CSU](#)
 - [Dryad](#) is a non-profit FAIR-aligned data repository
 - Data preserved in CoreTrustSeal-certified repository
 - Review and curation of all data sets
 - DataCite DOIs registered for all data sets
 - Free to deposit up to 300 GB per data set for CSU affiliated researchers
 - Requires ORCID for login
 - Over 400 CSU data sets published in Dryad

Top Considerations for You # 2:

Select a FAIR-aligned data repository

- General repositories (e.g., [Dryad](#), [Dataverse](#), [Zenodo](#))
 - Open to anyone to deposit
 - Minimal review/curation of deposits
 - Typically provide DataCite DOIs and usage metrics
 - Size limits and/or additional fees for large data
- Domain repositories:
 - [Re3data](#) repository registry
 - Level of review/curation varies
 - Ability to deposit varies
 - May be recommended/required for certain data types by funders/publishers (e.g., [Springer-Nature's list](#))



Top Considerations for You # 3:

Consider copyright and licensing of your data set

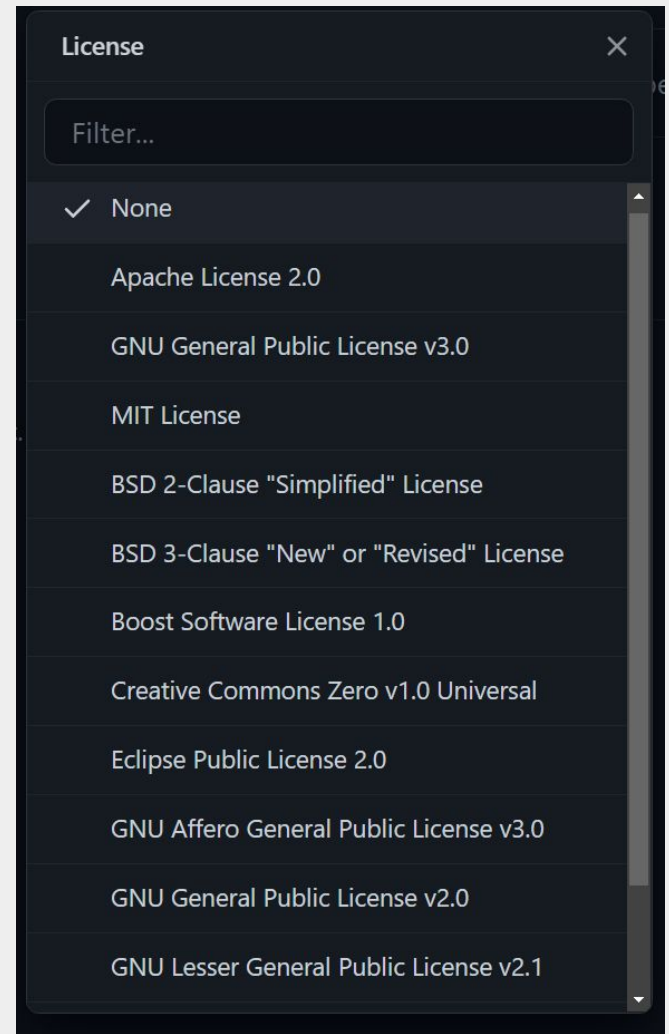
- The license that is selected facilitates sharing and reuse of the data set
 - [Creative Commons](https://creativecommons.org/licenses/by/4.0/)
 - CC BY: Creative Commons Attributions License
 - CC 0: When an owner wishes to waive their copyright and/or database rights
 - Public Domain mark (PDM): It is used to mark works that are in the public domain, and for which there are no known copyright or database restrictions.

Top Considerations for You # 3:

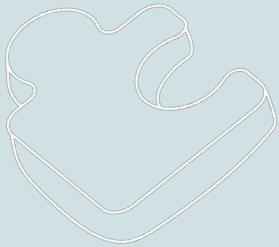
Consider copyright and licensing of associated software/code

- Many licenses available for software/code


GitHub



CU Scholar/Dryad Examples



Example Data Set in CU Scholar

 University of Colorado Boulder

CU Scholar
UNIVERSITY LIBRARIES

English Login


Home About Help Contact

Search CU Scholar

Enter search terms or select 'Go' to browse

Go

Home / Back to search results


 Data Set

Thermal Structure of the Martian Upper Mesosphere/Lower Thermosphere from MAVEN/IUVS Stellar Occultations [Data]

Public Deposited

Analytics

Citeable URL: <https://scholar.colorado.edu/concern/datasets/h702q775d>



Download

Citations Citations: ▾

Abstract

The MAVEN/IUVS stellar occultation dataset is publicly available at the NASA planetary data system's atmosphere node. But it consists of only the nightside events and limited dayside observations. We have reprocessed the mission-wide dataset from March 2015 to January 2022 using an improved stray light removal algorithm to retrieve the dayside events as well, thereby expanding the usable stellar occultation dataset and enabling the study of diurnal thermal structure of the upper mesosphere/lower thermosphere (~80-160 km).

We have provided this reprocessed dataset and the retrieved data products according to the executed campaigns, along with the output of a global-mean 1-D numerical model.


Creator


Gupta, Sumedha

Academic Affiliation

<https://doi.org/10.25810/z1wy-cq62>


Example Data Set in Dryad

 **DRYAD**

Explore data | 

Who we are | What we do | Join us | Help ▾ | Login


New indicators of ecological resilience and invasion resistance to support prioritization and management in the sagebrush biome, United States

Chambers, Jeanne, Rocky Mountain Research Station
Brown, Jessi, Rocky Mountain Research Station,  <https://orcid.org/0000-0002-6126-7951>
Bradford, John, United States Geological Survey
Board, David, Rocky Mountain Research Station
Campbell, Steven, Natural Resources Conservation Service
Clause, Karen, Natural Resources Conservation Service
Hanberry, Brice, Rocky Mountain Research Station
Schlaepfer, Daniel, United States Geological Survey
Urza, Alexandra, Rocky Mountain Research Station

jeanne.chambers@usda.gov, jessibrown@gmail.com, jbradford@usgs.gov,
david.i.board@usda.gov, brice.hanberry@usda.gov, dschlaepfer@usgs.gov,
alexandra.urza@usda.gov

Published Apr 13, 2023 on Dryad. <https://doi.org/10.5061/dryad.h18931zpb>

Data files

 Download dataset

> Jan 05, 2023
> Apr 13, 2023







Related works

Primary article
<https://doi.org/10.33...89/fevo.2022.1009268>


Software
<https://doi.org/10.5281/zenodo.7686426>

Supplemental information
<https://doi.org/10.5281/zenodo.7686427>

Share

Metrics

Cite this dataset 

Chambers, Jeanne et al. (2023). New indicators of ecological resilience and invasion resistance to support prioritization and management in the sagebrush biome, United States [Dataset]. Dryad.
<https://doi.org/10.5061/dryad.h18931zpb>

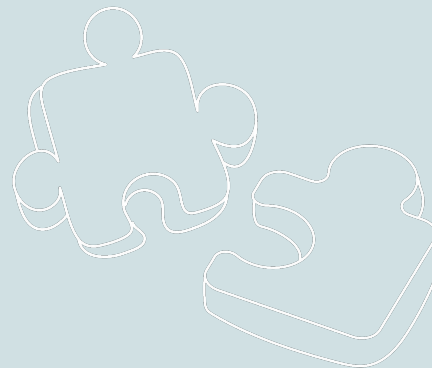
<https://doi.org/10.5061/dryad.h18931zpb>

CSU & Dryad Resources

- CSU Libraries Data Management website:
<https://lib.colostate.edu/services/data-management/dryad>
- Dryad best practices guide:
https://datadryad.org/stash/best_practices
- CSU Open Data guide:
<https://libguides.colostate.edu/openaccess/opendata>

Thank you!

Comments? Questions?



General Questions:
crdds@colorado.edu

CU Scholar Questions:
cuscholaradmin@colorado.edu

CSU & Dryad Questions:
mara.sedlins@colostate.edu



Center for Research Data & Digital Scholarship
UNIVERSITY OF COLORADO **BOULDER**