

Research Data Camp: Data Publishing and Repositories



Melissa Cantrell, Assistant Professor, Scholarly Communication Librarian

Matthew Murray, Teaching Assistant Professor, Data Librarian

August 20th, 2024



Center for Research Data & Digital Scholarship

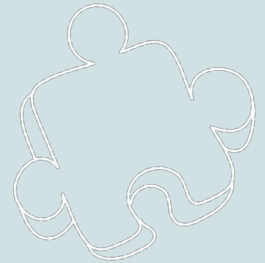
UNIVERSITY OF COLORADO **BOULDER**

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

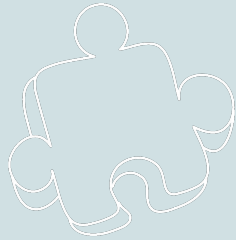


**Open, accessible, and
reproducible data
advancing the public good**

**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



ACCESIBLE

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)

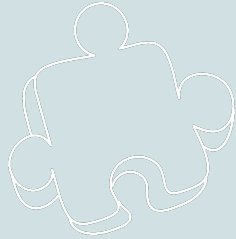
- Origin, context, history, and who to credit/cite are all crucial for data reusability
- Consider permitted use and apply the appropriate license

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

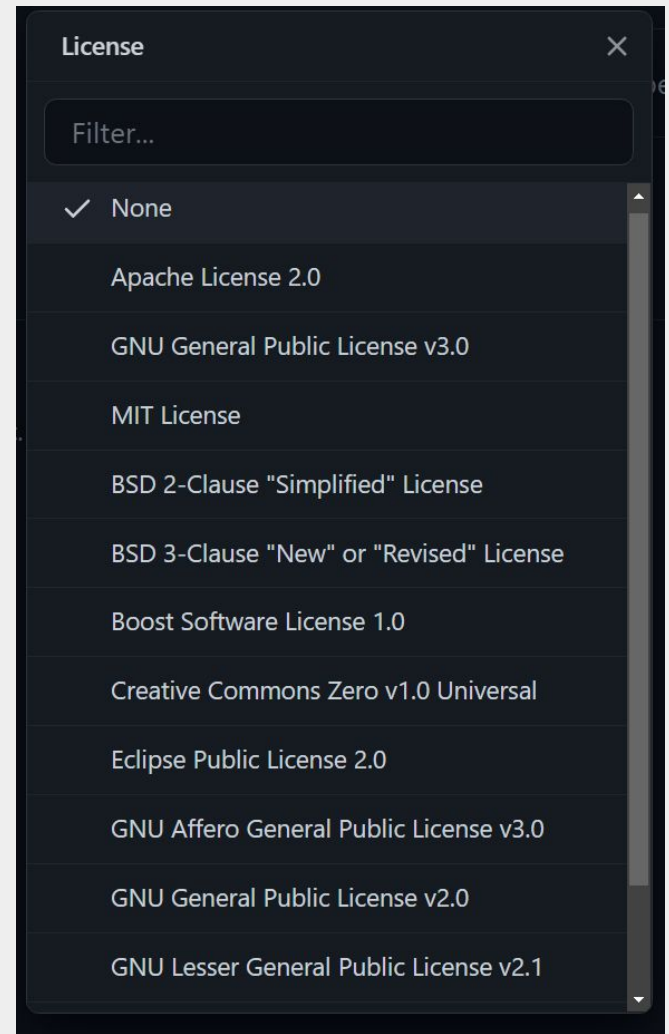
- The license that is selected facilitates sharing and reuse of the data set
 - [Creative Commons](#)
 - CC BY: Creative Commons Attributions License
 - CC 0: When an owner wishes to waive their copyright and/or database rights
 - Public Domain mark (PDM): It is used to mark works that are in the public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

Top Considerations for You # 3:

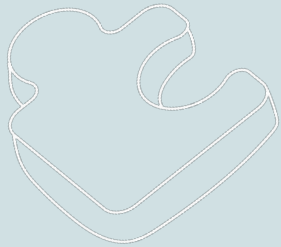
Consider copyright and licensing of associated software/code

- 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

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>

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
Laboratory for Atmospheric & Space Physics

Last Modified
2022-08-12

Resource Type
Data Set

Rights Statement
In Copyright [↗](#)

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








Language
English (eng) [↗](#)

Citation
Gupta, S. (2022). Thermal Structure of the Martian Upper Mesosphere/Lower Thermosphere from MAVEN/IUVS Stellar Occultations [Data] [Data set]. University of Colorado Boulder. <https://doi.org/10.25810/z1wy-cq62>

License
Creative Commons BY Attribution 4.0 International [↗](#)

Relationships

Items

Thumbnail	Title	Date Uploaded	Visibility	Actions
	README.pdf	2022-08-09	Public	Download
	1-D_model.xlsx	2022-08-09	Public	Download
	Campaign_2.zip	2022-08-09	Public	Download
	Campaign_3.zip	2022-08-09	Public	Download
	Campaign_5.zip	2022-08-09	Public	Download
	Campaign_6.zip	2022-08-09	Public	Download
	Campaign_7.zip	2022-08-09	Public	Download

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

HOME

ABOUT

HELP

CONTACT

Search CU Scholar

Enter search terms or select 'Go' to browse

Go

All

Share Your Work

Terms of Use

Featured Works

Recently Uploaded



Genetic and Environmental Etiologies of Reading Disabilities: Analysis of data from the Colorado Learning Disabilities Research Center

Creator: Astrom, Raven

Subject: Twins, Heritability, Longitudinal, Reading, Disability, Siblings

Resource Type: Dissertation



The 19 Percent: Disability and Actor Training in Higher Education

Creator: McNish, Deric

Subject: Universal Design for Learning, Performance, Disability, Voice, Actor Training

Resource Type: Dissertation



Exploring Web Simplification for People with Cognitive Disabilities

Creator: Hoehl, Jeffrey Arthur

Subject: human computer computing, web simplification, cognitive disabilities, accessibility, human computer interaction, web accessibility

Resource Type: Dissertation



Save order

Explore Collections

Featured Researcher



Works



Electromagnetics Laboratory/The MIMICAD Research Center



Series in Biology



University Libraries Fellows



Western States Government Information Virtual Conference

View all collections



Terms of Use

Featured Works Recently Uploaded

☰

Genetic and Environmental Etiologies of Reading Disabilities: Analysis of data from the Colorado Learning Disabilities Research Center ✕

Creator: Astrom, Raven

Subject: Twins, Heritability, Longitudinal, Reading, Disability, Siblings

Resource Type: Dissertation

☰

The 19 Percent: Disability and Actor Training in Higher Education ✕

Creator: McNish, Deric

Subject: Universal Design for Learning, Performance, Disability, Voice, Actor Training

Resource Type: Dissertation

☰

Exploring Web Simplification for People with Cognitive Disabilities ✕

Creator: Hoehl, Jeffrey Arthur

Subject: human computer computing, web simplification, cognitive disabilities, accessibility, human computer interaction, web accessibility

Resource Type: Dissertation

Save order

Explore Collections Featured Researcher

Works

Electromagnetics Laboratory/The MIMICAD Research Center

Series in Biology

University Libraries Fellows

Western States Government Information Virtual Conference

View all collections

Featured Works Recently Uploaded

- Genetic and Environmental Analysis of data from the Center
Creator: Astrom, Raven
Subject: Twins, Heritability
Resource Type: Dissertation
- The 19 Percent: Disability
Creator: McNish, Deric
Subject: Universal Design, Training
Resource Type: Dissertation
- Exploring Web Simplification
Creator: Hoehl, Jeffrey A
Subject: human computer interaction, accessibility, human computer interaction
Resource Type: Dissertation

Save order

Select type of work

- Graduate Thesis or Dissertation
Graduate theses or dissertations
- Undergraduate Honors Thesis
Undergraduate honors theses
- Article
Research articles, reviews, editorials, etc.
- Data Set
Research data sets
- Presentation
Presentations
- Conference Proceeding
Conference proceedings
- Book
Books
- Book Chapter
Book chapters
- Report
Technical reports, working papers, reports, etc.
- Other Works
Digital content that does not fit into any of these other categories

Close Create work

to browse Go All

MIMICAD Research Center

Information Virtual Conference

Matthew Murray

Home / Dashboard / Works / Add New Work

- ACTIVITY
 - Your activity
 - Reports
- REPOSITORY CONTENTS
 - Collections
 - Works
- TASKS
 - Review Submissions
 - Manage Embargoes
 - Manage Leases
- CONFIGURATION
 - Settings
 - Workflow Roles

Add New Data Set

Descriptions Files Relationships

To create a separate work for each of the files, go to [Batch upload](#)

Title required

A name to aid in identifying a work.

[+ Add another Title](#)

Creator required

The person or group responsible for the work. Usually this is the author of the content. Personal names should be entered with the last name first, e.g. "Smith, John."

[+ Add another Creator](#)

Academic Affiliation required

Academic Department/College/School/Unit at CU Boulder

[+ Add another Academic Affiliation](#)

Resource Type required

The general category of this resource (e.g. Masters Thesis, dissertation).

Rights Statement required

[Additional fields](#)

Save Work

Requirements

- Describe your work
- Add files
- Check deposit agreement

Visibility

- Public**
Make available to all.
Please note, making something visible to the world (i.e. marking this as **Public**) may be viewed as publishing which could impact your ability to:
 - Patent your work
 - Publish your work in a journal
 Check out [SHERPA/RoMEO](#) for more information about publisher copyright policies.
- Embargo**
Set date for future release.

I have read and agree to the [Deposit Agreement](#)

Save

[Additional fields](#)

Date Issued

The date the resource was published or awarded, such as when an article is published in a journal. Format: yyyy-mm-dd

Abstract or Summary

A brief description or summary of the item.

File Edit View Format

Formats B I

POWERED BY TINY

[X Remove](#)

Subject

Headings or index terms describing what the work is about.

[+ Add another Subject](#)

Example Data Set in 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 ¹ ; Brown, Jessi ¹ ; Bradford, John ² ; Board, David ¹ ; Campbell, Steven ² ; Clause, Karen ² ; Hanberry, Brice ¹ ; Schlaepfer, Daniel ² ; Urza, Alexandra ¹ 

Author affiliations 

Published Jan 05, 2023; Updated Apr 13, 2023 on Dryad. <https://doi.org/10.5061/dryad.h18931zpb>

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>

Abstract

Ecosystem transformations to altered or novel ecological states are accelerating across the globe. Indicators of ecological resilience to disturbance and resistance to invasion can aid in assessing risks and prioritizing areas for conservation and restoration. The sagebrush biome encompasses parts of 11 western states and is experiencing rapid transformations due to human population growth, invasive species, altered disturbance regimes, and climate change. We built on prior use of static soil moisture and temperature regimes to develop new, ecologically relevant and climate-responsive indicators of both resilience and resistance. Our new indicators were based on climate and soil water availability variables derived from process-based ecohydrological models that allow predictions of future conditions. We asked: (1) Which variables best indicate resilience and resistance? (2) What are the relationships among the indicator variables and resilience and resistance categories? (3) How do patterns of resilience and resistance vary across the area? We assembled a large database (n = 24,045) of vegetation sample plots from regional monitoring programs and derived multiple climate and soil water availability variables for each plot from ecohydrological simulations. We used USDA Natural Resources Conservation Service National Soils Survey Information, Ecological Site Descriptions, and expert knowledge to develop and assign ecological types and resilience and resistance categories to each plot. We used random forest models to derive a set of 19 climate and water availability variables that best predicted resilience and resistance categories. Our models had relatively high multiclass accuracy (80% for resilience; 75% for resistance). Top indicator variables for both resilience and resistance included mean temperature, coldest month temperature, climatic water deficit, and summer and driest month precipitation. Variable relationships and patterns differed among ecoregions but

Data files

> Jan 05, 2023 version files	294.65 MB
✓ Apr 13, 2023 version files	1.36 GB
Data.zip	1.35 GB
README.html	663.66 KB
README.md	33.59 KB

 Download full dataset

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

 461 views

 135 downloads

 4 citations

resilient environmental gradients, low resilience and resistance were indicated by warm and dry conditions with high climatic water deficits, and moderately high to high resilience and resistance were characterized by cooler and moister conditions with low climatic water deficits. The new, ecologically-relevant indicators provide information on the vulnerability of resources and likely success of management actions and can be used to develop new approaches and tools for prioritizing areas for conservation and restoration actions.

Methods

We assembled a large database (n = 24,045) of vegetation sample plots from regional monitoring programs and derived multiple climate and soil water availability variables for each plot from ecohydrological simulations. We used USDA Natural Resources Conservation Service National Soils Survey Information, Ecological Site Descriptions, and expert knowledge to develop and assign ecological types and resilience and resistance categories to each plot. We used random forest models to derive a set of 19 climate and water availability variables that best predicted resilience and resistance categories.

Usage notes

All code scripts are RStudio notebooks, which are RMarkdown files additionally formatted to render to HTML. Input files can be .csv files (plain text, common-separated files) or RDS files (R data objects).

Funding

Joint Fire Sciences Program, Award: Project 19-2-02-11

Rocky Mountain Research Station

Subject keywords

Natural sciences
Bromus tectorum
Climate
ecohydrological simulation
ecological resilience
prioritization
random forest
resistance to invasion
sagebrush biome
Soil water availability

License

This work is licensed under a [CC0 1.0 Universal \(CC0 1.0\) Public Domain Dedication](https://creativecommons.org/licenses/by/4.0/) license.

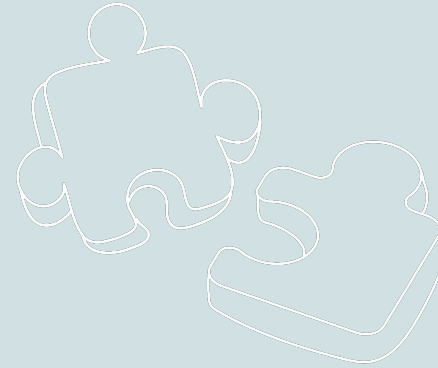


CSU & Dryad Resources

- 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**