Tales from the AGU Data Help Desk:

Data/Software Availability Statements, Citations, and More

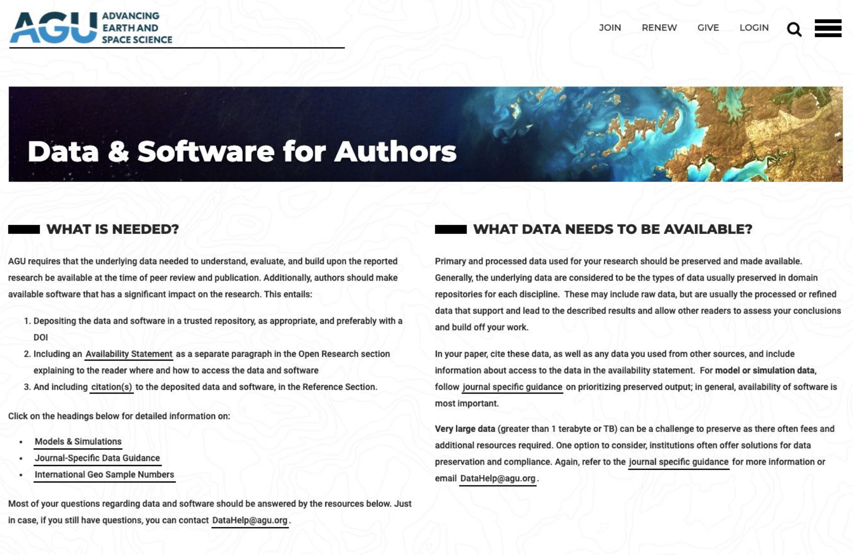
21 April 2022

Chris Erdmann
Assistant Director, Data Leadership
American Geophysical Union
0000-0003-2554-180X
@libcce | cerdmann@agu.org

AGU Data & Software Sharing Guidance

What is covered:

- What data needs to be available?
- Repository Selection
- Availability Statement
- Data & Software Citation
- Citation Formatter
- Models & Simulations
- Journal Specific Guidance
- International Geo Sample Numbers
- **Data Help Desk**



Data & Software for Authors

WHAT IS NEEDED?

AGU requires that the underlying data needed to understand, evaluate, and build upon the reported research be available at the time of peer review and publication. Additionally, authors should make available software that has a significant impact on the research. This entails:

1. Depositing the data and software in a trusted repository, as appropriate, and preferably with a DOI
2. Including an Availability Statement as a separate paragraph in the Open Research section explaining to the reader where and how to access the data and software
3. And including citation(s) to the deposited data and software, in the Reference Section.

Click on the headings below for detailed information on:

- Models & Simulations
- Journal-Specific Data Guidance
- International Geo Sample Numbers

Most of your questions regarding data and software should be answered by the resources below. Just in case, if you still have questions, you can contact DataHelp@agu.org.

WHAT DATA NEEDS TO BE AVAILABLE?

Primary and processed data used for your research should be preserved and made available. Generally, the underlying data are considered to be the types of data usually preserved in domain repositories for each discipline. These may include raw data, but are usually the processed or refined data that support and lead to the described results and allow other readers to assess your conclusions and build off your work.

In your paper, cite these data, as well as any data you used from other sources, and include information about access to the data in the availability statement. For **model or simulation data,** follow journal specific guidance on prioritizing preserved output; in general, availability of software is most important.

Very large data (greater than 1 terabyte or TB) can be a challenge to preserve as there often fees and additional resources required. One option to consider, institutions often offer solutions for data preservation and compliance. Again, refer to the journal specific guidance for more information or email DataHelp@agu.org.

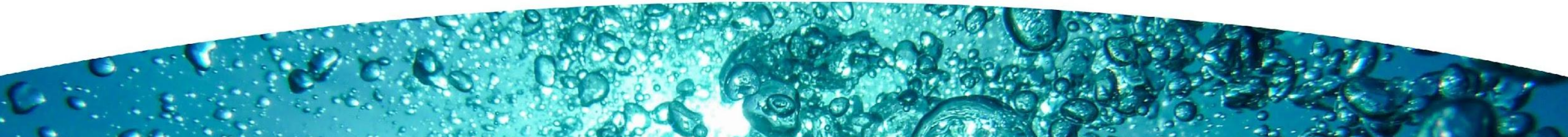
https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Data-and-Software-for-Authors

Main Goal: Avoid parachuting researchers into your data/software and instead guide them as best as you can

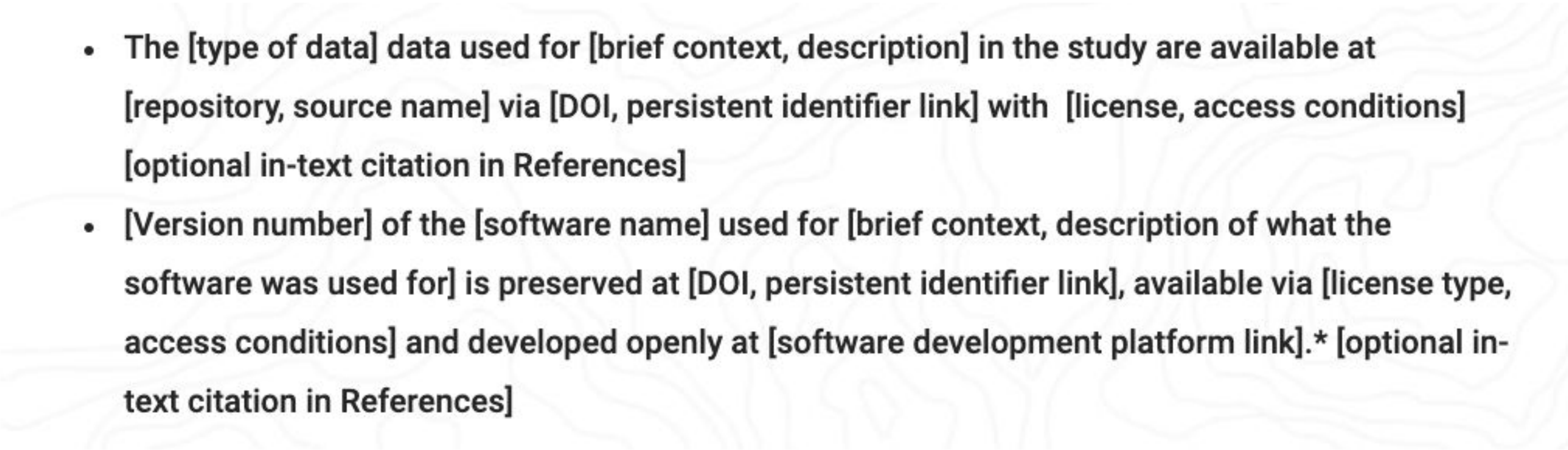
[Learning to get from A to B, one windy afternoon in the New Forest]() by Annie Spratt
[Paragliding above the Chartreuse massif]() by Nicolas Tissot

What is included in an Availability Statement?

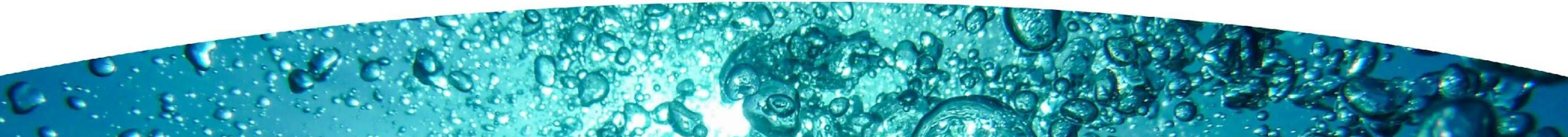
1. A brief description of the type(s) of data or software
2. Repository Name(s) where they are deposited
3. Version (of software)
4. **DOI, Persistent Identifier** Link to Data or Software
5. Link to publicly accessible development platform (in the case of Software, e.g. GitHub)
6. Access Conditions (e.g. if Registration is Required)
7. Licensing/Permissions (e.g. Creative Commons Attribution)
8. **In-text citation in References (optional)**



Availability Statement Templates

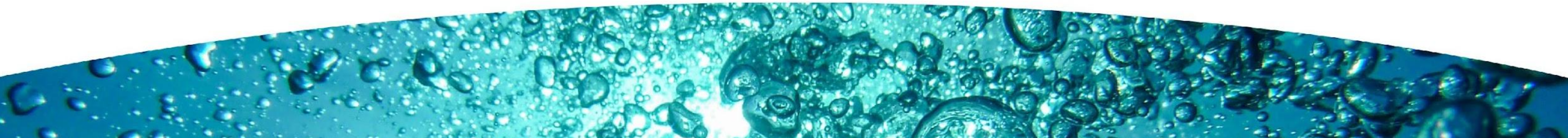


- The [type of data] data used for [brief context, description] in the study are available at [repository, source name] via [DOI, persistent identifier link] with [license, access conditions] [optional in-text citation in References]
- [Version number] of the [software name] used for [brief context, description of what the software was used for] is preserved at [DOI, persistent identifier link], available via [license type, access conditions] and developed openly at [software development platform link].* [optional in-text citation in References]



What is included in a data/software citation?

1. Author(s) or project name(s)
2. Date / Software published
3. Title / Software name
4. Data or software release/version (optional)
5. **Bracketed description type (e.g., [Dataset], [Software], [Collection], [ComputationalNotebook])**
6. Repository name / Publication venue
7. **DOI, persistent identifier, URL**
8. Retrieved date (required when using URL)



Data Citation Examples

- Fiechter, J., & Cheresh, J. (2020). Physical and biogeochemical drivers of alongshore pH and oxygen variability in the California Current System (Version 7) [Dataset]. Dryad. https://doi.org/10.7291/D1D96Q

- Edmunds, P. J., Didden, C., & Frank, K. (2021). Mean percentage cover of corals and Porites astreoides at each site by year at St. John, VI from 1992 to 2019 (Version 1) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/10.26008/1912/BCO-DMO.843284.1

- Alwarda, R., & Smith, I. (2021). Elevation data for Reflectors within the CO2 Deposit in Planum Australe, Mars [Dataset]. Zenodo. https://doi.org/10.5281/ZENODO.4639669

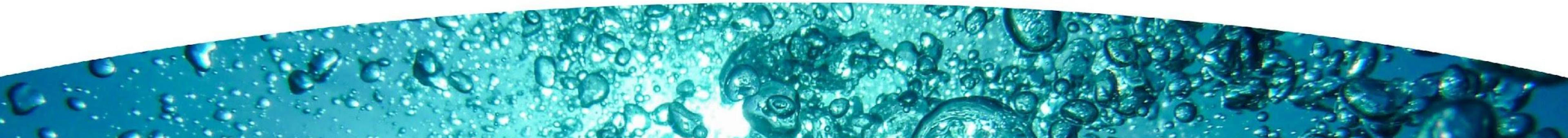
- Gries, C., Downs, R. R., O'Brien, M., Parr, C., Duerr, R., Koskela, R., et al. (2019). Return on Investment Metrics for Data Repositories in Earth and Environmental Sciences [Dataset]. Environmental Data Initiative. https://doi.org/10.6073/PASTA/D49BEC63F51603512EFA7E0FD2717203

Software Citation Examples

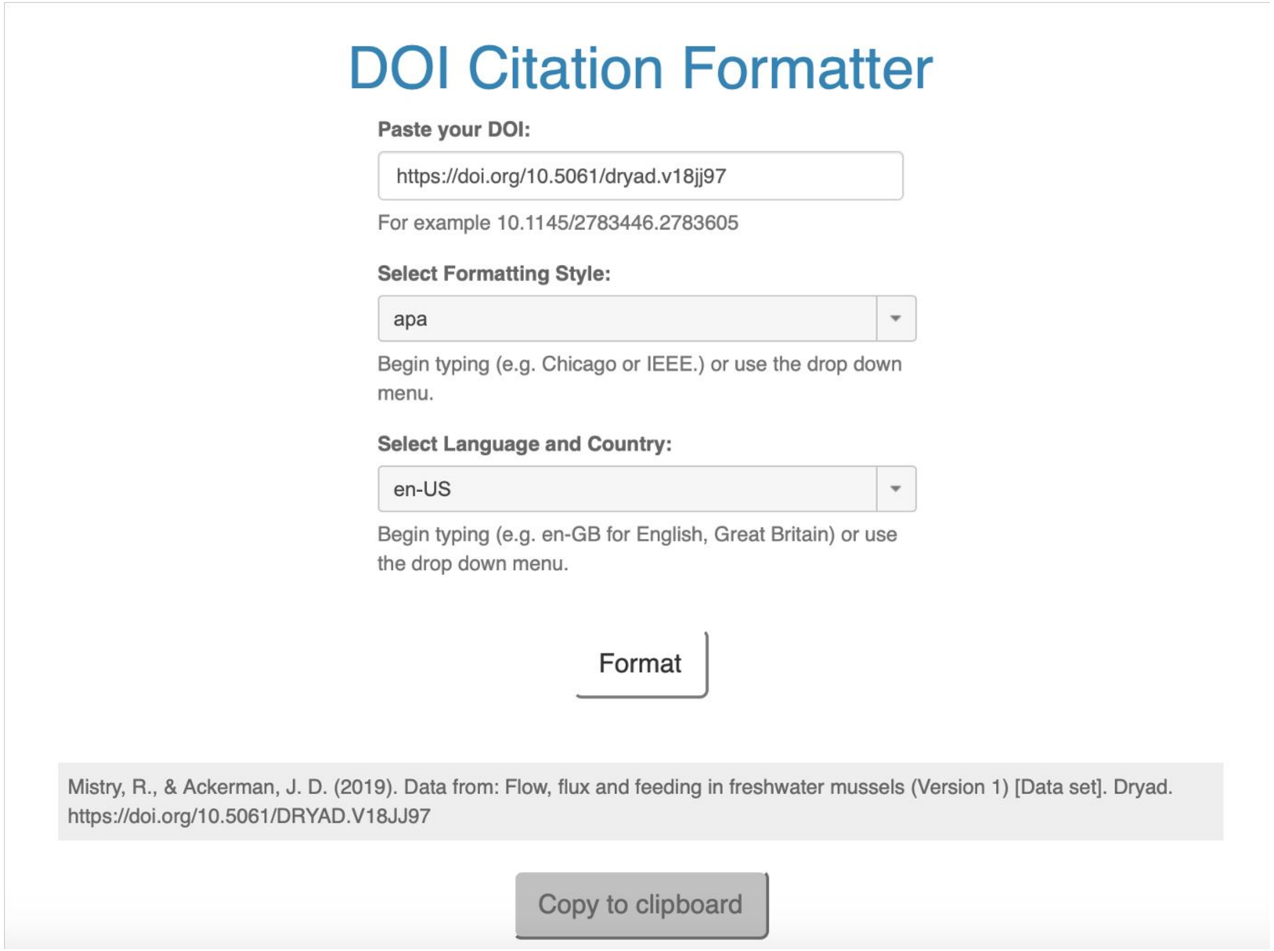
- Lab for Exosphere and Near Space Environment Studies. (2019, March 20). lenses-lab/LYAO_RT-2018JA026426: Original Release (Version 1.0.0) [Software]. Zenodo. http://doi.org/10.5281/zenodo.2598836

- Bell, S. W. (2020). samwbell/saturn_counts: April 26, 2020 Release (Version 1.1.0) [Software]. Zenodo. https://doi.org/10.5281/ZENODO.3766959

- Shaoqian Hu. (2019, December 25). Direct surface wave radial anisotropy tomography package (Version 1.0) [Software]. Zenodo. http://doi.org/10.5281/zenodo.3592528



AGU
ADVANCING EARTH
AND SPACE SCIENCE



DOI Citation Formatter

Paste your DOI:

https://doi.org/10.5061/dryad.v18jj97

For example 10.1145/2783446.2783605

Select Formatting Style:

apa

Begin typing (e.g. Chicago or IEEE.) or use the drop down menu.

Select Language and Country:

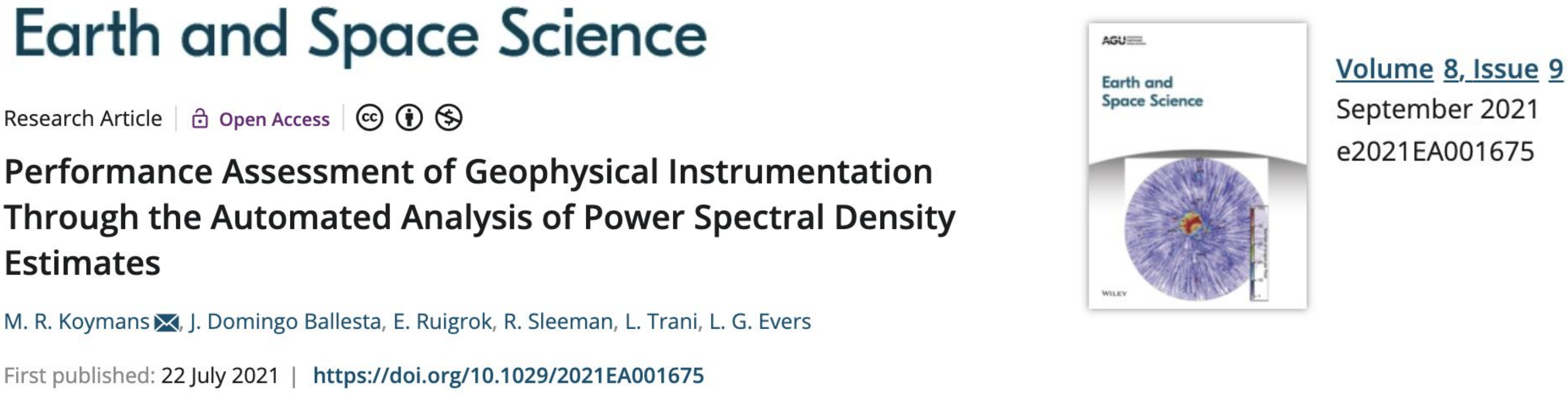
en-US

Begin typing (e.g. en-GB for English, Great Britain) or use the drop down menu.

Mistry, R., & Ackerman, J. D. (2019). Data from: Flow, flux and feeding in freshwater mussels (Version 1) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.V18JJ97

https://citation.crosscite.org

Availability Statement/Citation Paper Example



Earth and Space Science

Research Article | Open Access | 🅭 🅯 🄏

Performance Assessment of Geophysical Instrumentation Through the Automated Analysis of Power Spectral Density Estimates

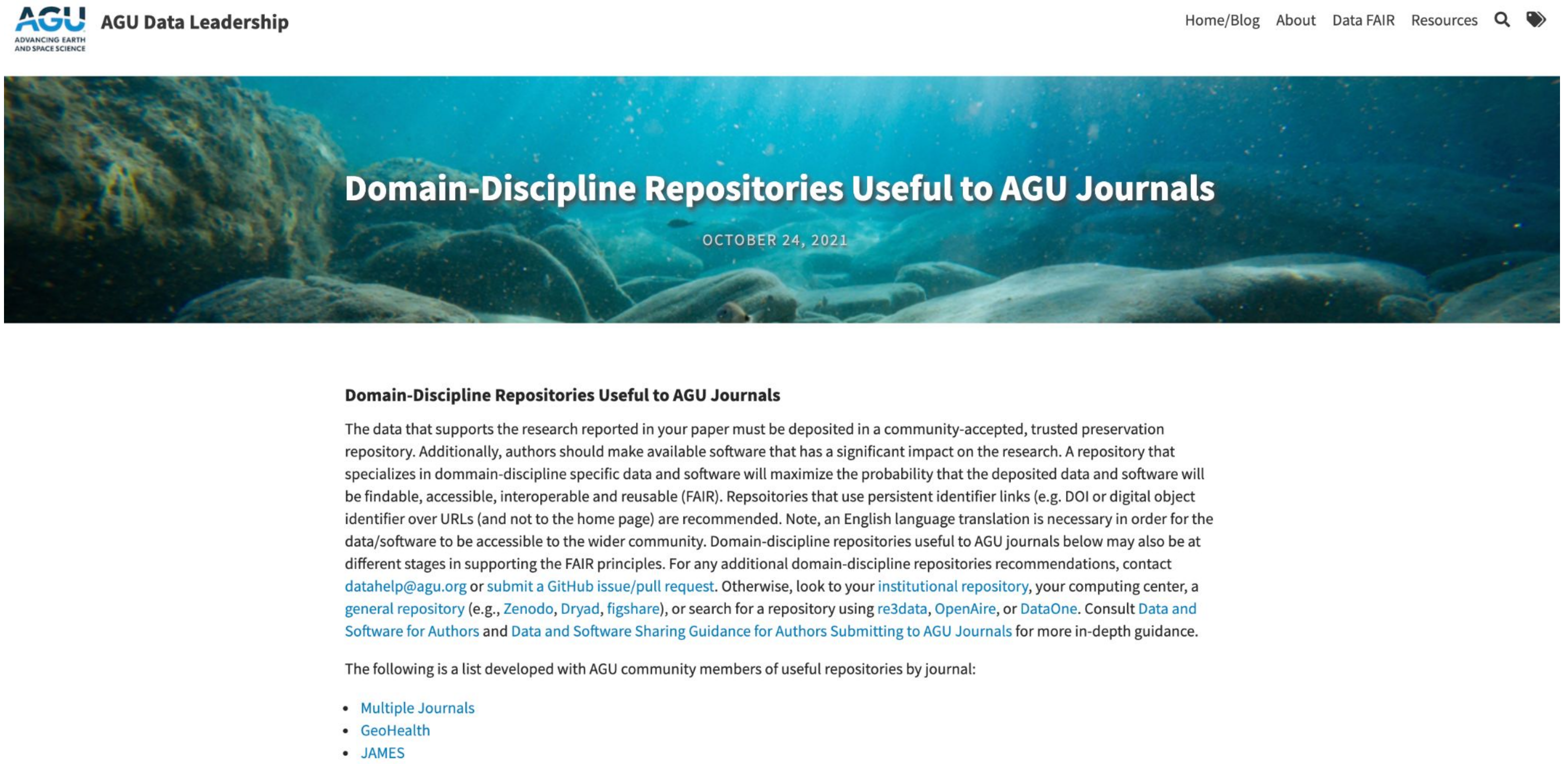
M. R. Koymans ✉, J. Domingo Ballesta, E. Ruigrok, R. Sleeman, L. Trani, L. G. Evers

First published: 22 July 2021 | https://doi.org/10.1029/2021EA001675

Volume 8, Issue 9
September 2021
e2021EA001675

https://doi.org/10.1029/2021EA001675

What repository?



Domain-Discipline Repositories Useful to AGU Journals

OCTOBER 24, 2021

Domain-Discipline Repositories Useful to AGU Journals

The data that supports the research reported in your paper must be deposited in a community-accepted, trusted preservation repository. Additionally, authors should make available software that has a significant impact on the research. A repository that specializes in dommain-discipline specific data and software will maximize the probability that the deposited data and software will be findable, accessible, interoperable and reusable (FAIR). Repsoitories that use persistent identifier links (e.g. DOI or digital object identifier over URLs (and not to the home page) are recommended. Note, an English language translation is necessary in order for the data/software to be accessible to the wider community. Domain-discipline repositories useful to AGU journals below may also be at different stages in supporting the FAIR principles. For any additional domain-discipline repositories recommendations, contact datahelp@agu.org or submit a GitHub issue/pull request. Otherwise, look to your institutional repository, your computing center, a general repository (e.g., Zenodo, Dryad, figshare), or search for a repository using re3data, OpenAire, or DataOne. Consult Data and Software for Authors and Data and Software Sharing Guidance for Authors Submitting to AGU Journals for more in-depth guidance.

The following is a list developed with AGU community members of useful repositories by journal:

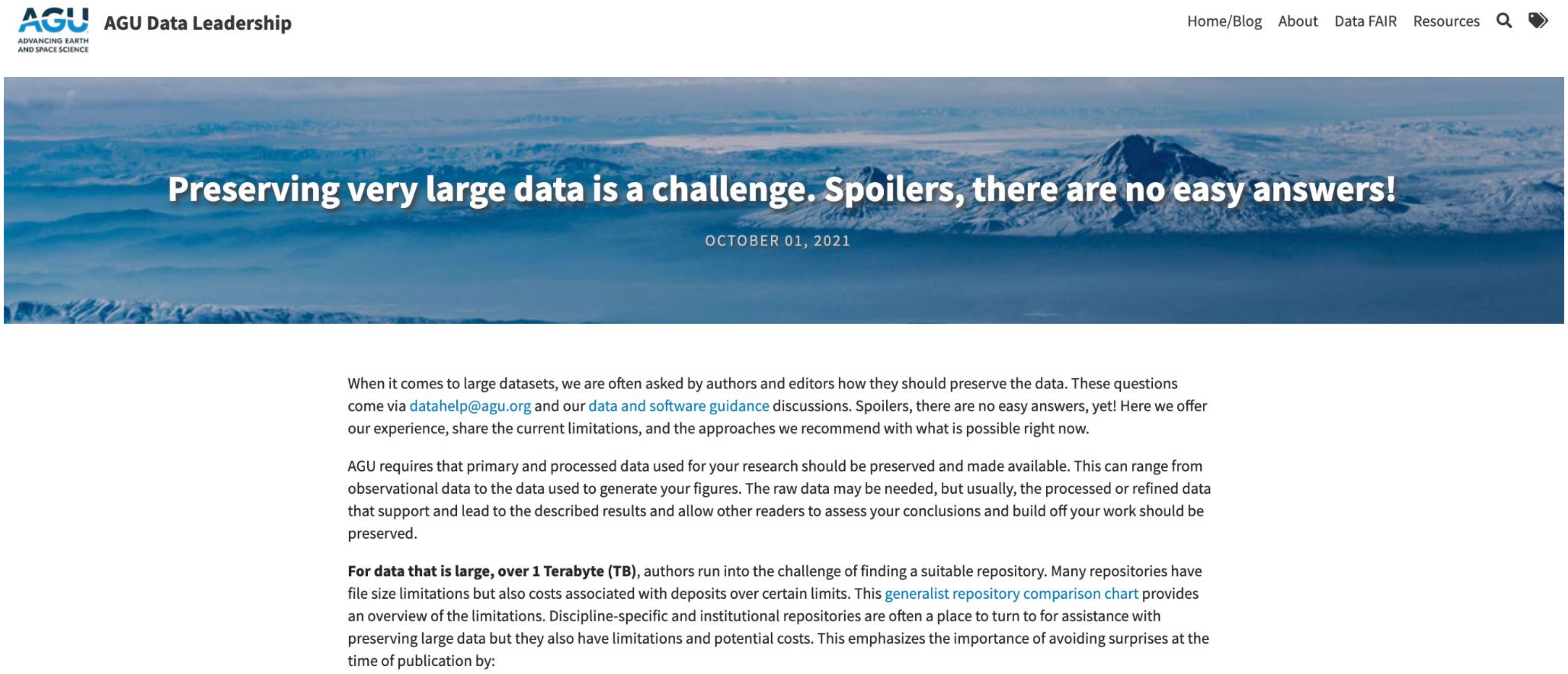
- Multiple Journals
- GeoHealth
- JAMES

https://data.agu.org/resources/useful-domain-repositories

Help Desk Challenges

- Government Sites, Similar - Technical, Permissions
- Firewalls, Authentication - Openness, Availability, Anonymity
- Supplemental Information - Tradition, Peer Review
- FTP, Directories, Storage - Institutional, Compliant Solution
- Curation, Deposit Workflows - Service, Publication Workflows
- Web Sharing, Dev Platforms - Citation Information
- Databases / Dynamic Services - Direct Access, Linking
- Available Upon Request - Culture
- Citation Nothingness (Paper not the Data) - Culture
- Website Home (Parachuting) - Laziness
- English Language - Language Diversity, Translation
- Many Data Links/Citations - Tables, Supplements (See Data Citation Community of Practice)
- ...

Preserving Large Data!



Preserving very large data is a challenge. Spoilers, there are no easy answers!

OCTOBER 01, 2021

When it comes to large datasets, we are often asked by authors and editors how they should preserve the data. These questions come via datahelp@agu.org and our data and software guidance discussions. Spoilers, there are no easy answers, yet! Here we offer our experience, share the current limitations, and the approaches we recommend with what is possible right now.

AGU requires that primary and processed data used for your research should be preserved and made available. This can range from observational data to the data used to generate your figures. The raw data may be needed, but usually, the processed or refined data that support and lead to the described results and allow other readers to assess your conclusions and build off your work should be preserved.

For data that is large, over 1 Terabyte (TB), authors run into the challenge of finding a suitable repository. Many repositories have file size limitations but also costs associated with deposits over certain limits. This generalist repository comparison chart provides an overview of the limitations. Discipline-specific and institutional repositories are often a place to turn to for assistance with preserving large data but they also have limitations and potential costs. This emphasizes the importance of avoiding surprises at the time of publication by:

DATA FAIR

And Software!

Get Cited, Get Funding and Go Further with Better Data and Software Skills

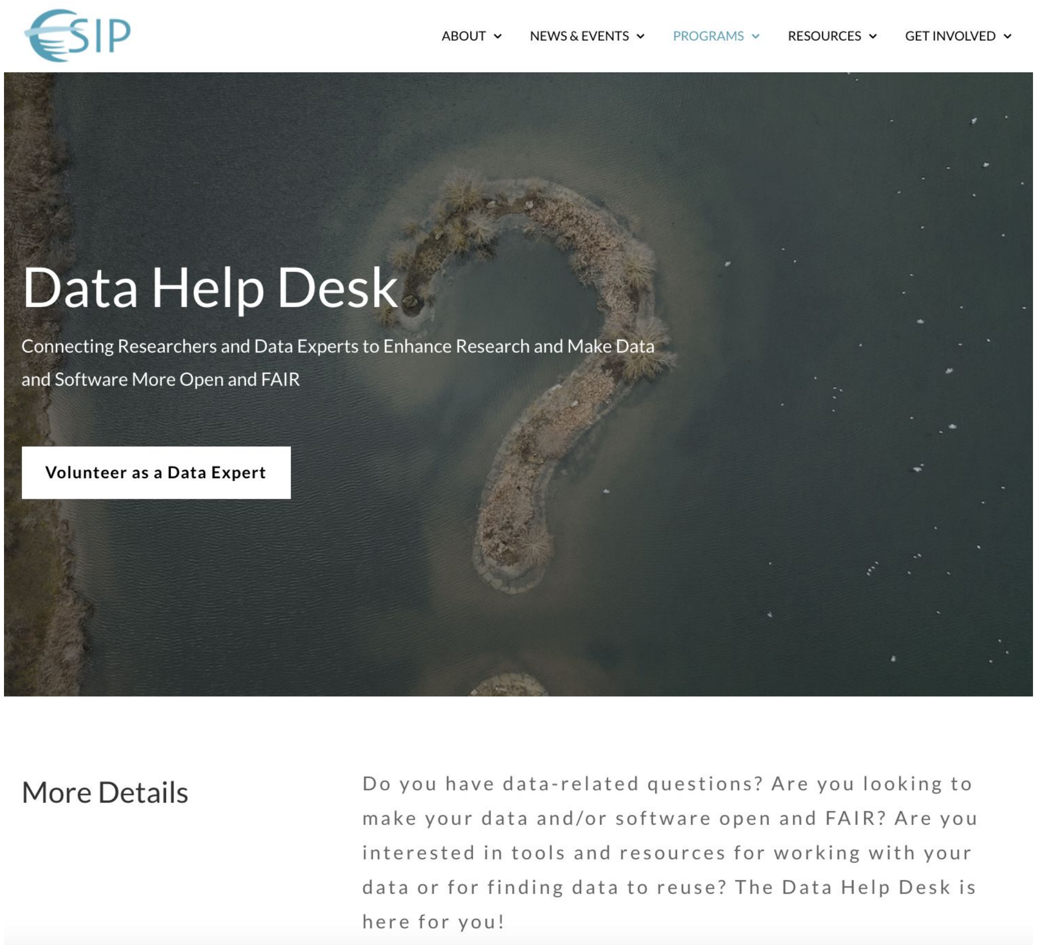
Data Help Desk • Data and Software Management Workshops • Tool and Platform Demos and Training

AGU Fall Meeting 2021

A program of ESIP, AGU, and their partners

ESIP

AGU ADVANCING EARTH AND SPACE SCIENCE



Data Help Desk

Connecting Researchers and Data Experts to Enhance Research and Make Data and Software More Open and FAIR

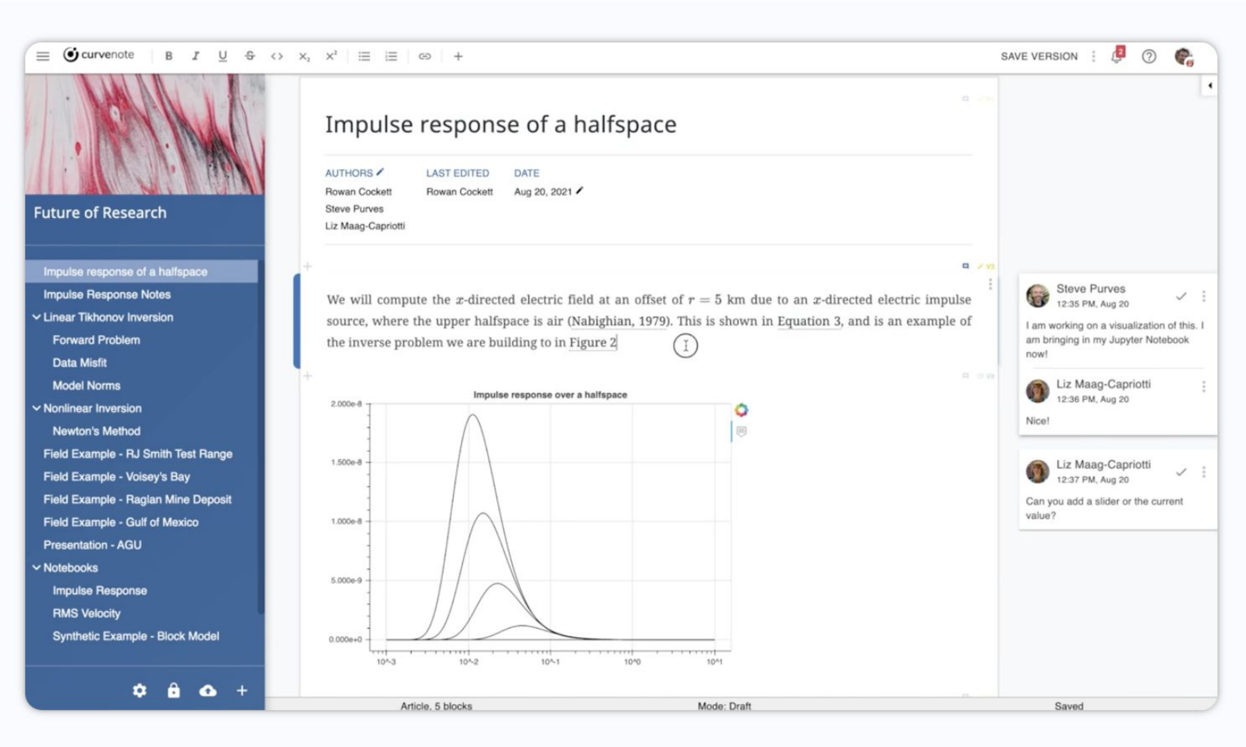
More Details

Do you have data-related questions? Are you looking to make your data and/or software open and FAIR? Are you interested in tools and resources for working with your data or for finding data to reuse? The Data Help Desk is here for you!

https://www.esipfed.org/data-help-desk-at-agu-2021

AGU
ADVANCING EARTH
AND SPACE SCIENCE

QC/Rules Before Manuscript Submission



Impulse response of a halfspace

AUTHORS
Rowan Cockett
Steve Purves
Liz Maag-Capriotti

LAST EDITED
Rowan Cockett

DATE
Aug 20, 2021

We will compute the x-directed electric field at an offset of $r = 5$ km due to an x-directed electric impulse source, where the upper halfspace is air (Nabighian, 1979). This is shown in Equation 3, and is an example of the inverse problem we are building to in Figure 2

Steve Purves
12:35 PM, Aug 20

I am working on a visualization of this. I am bringing in my Jupyter Notebook now!

Liz Maag-Capriotti
12:36 PM, Aug 20

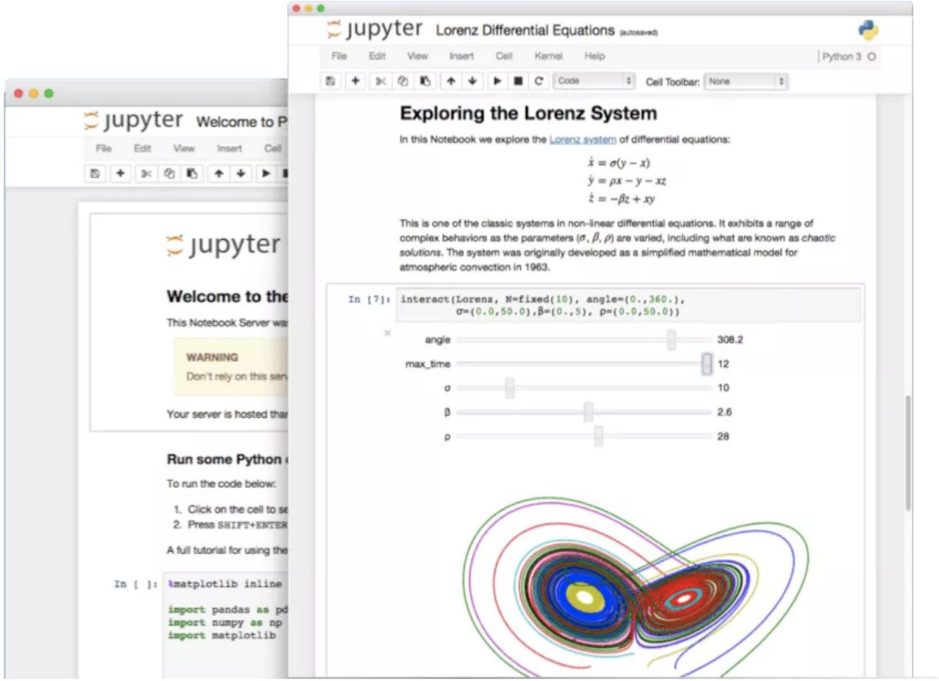
Nice!

Liz Maag-Capriotti
12:37 PM, Aug 20

Can you add a slider or the current value?

Example: https://curvenote.com/

Notebooks Now!



Code

Cell Toolbar: None

jupyter

Welcome to the

This Notebook Server wa

WARNING

Don't rely on this serv

Your server is hosted than

Run some Python

To run the code below:

1. Click on the cell to se
2. Press SHIFT+ENTER

A full tutorial for using the

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) are varied, including what are known as *chaotic solutions*. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.),
                 σ=(0.0,50.0),β=(0.,5), ρ=(0.0,50.0))
```

angle 308.2

max_time 12

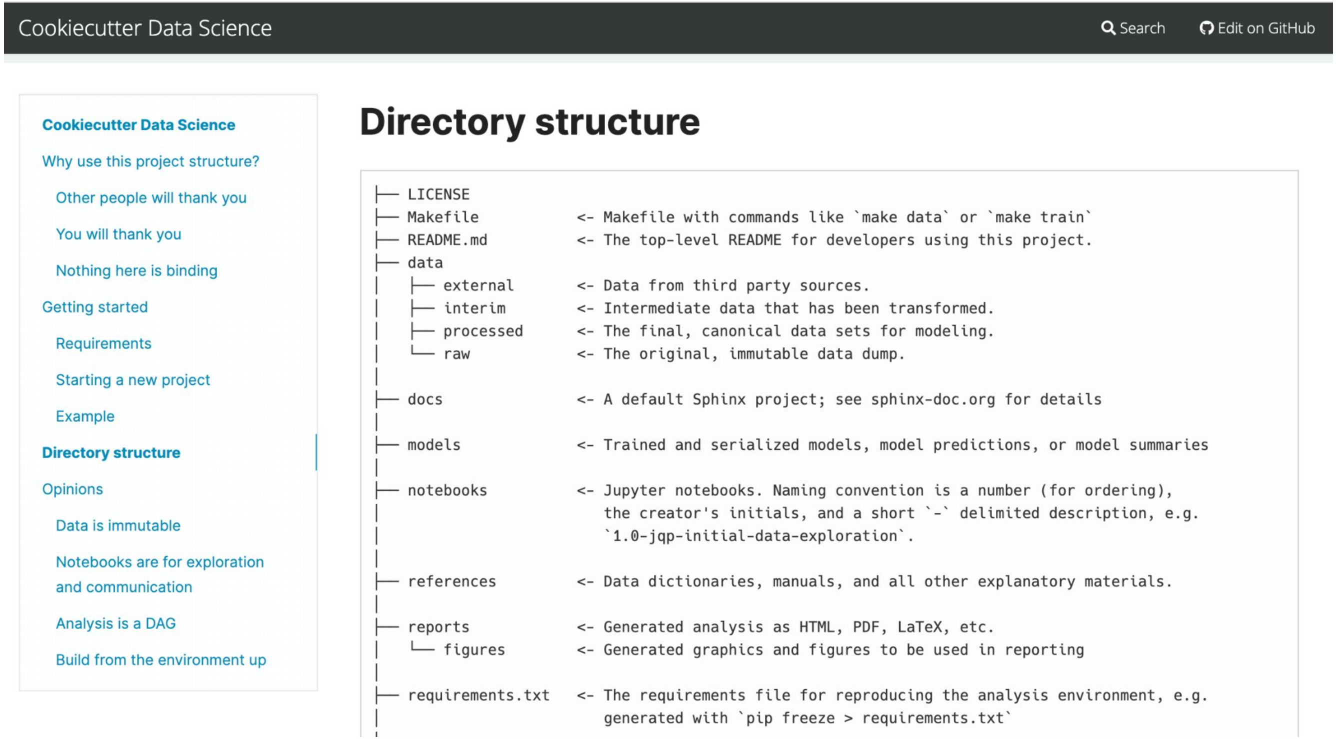
σ 10

β 2.6

ρ 28

Example: https://jupyter.org/

Cookiecutter Data Science



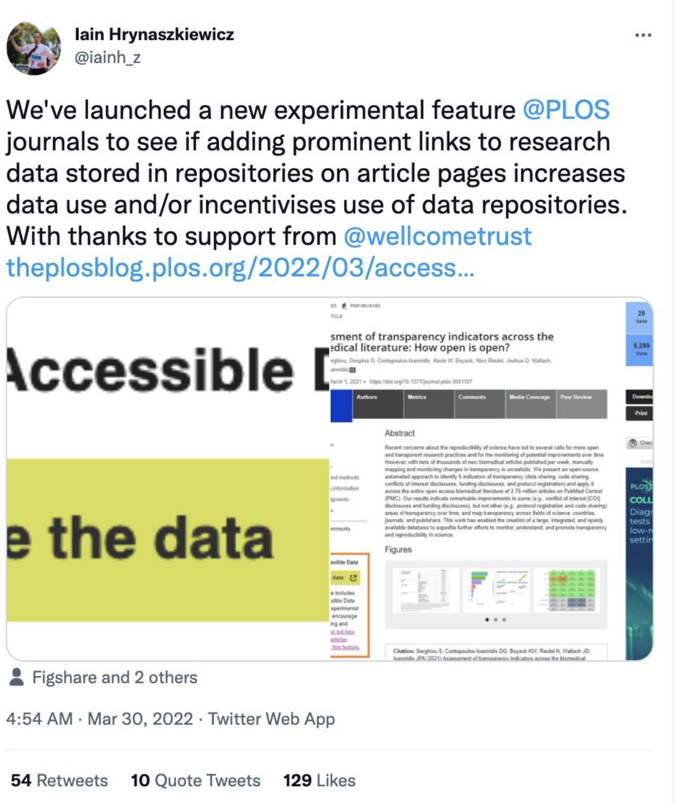
Cookiecutter Data Science

Directory structure

```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── requirements.txt   <- The requirements file for reproducing the analysis environment, e.g.
│                         generated with `pip freeze > requirements.txt`
│
```

https://drivendata.github.io/cookiecutter-data-science/

Indexing/Filtering



Iain Hrynaszkiewicz
@iainh_z

We've launched a new experimental feature @PLOS journals to see if adding prominent links to research data stored in repositories on article pages increases data use and/or incentivises use of data repositories. With thanks to support from @wellcometrust theplosblog.plos.org/2022/03/access...

Figshare and 2 others

4:54 AM · Mar 30, 2022 · Twitter Web App

54 Retweets **10** Quote Tweets **129** Likes

https://twitter.com/iainh_z/status/1509091657131638792

Takeaways

- We need everyone’s help advancing data/software sharing policies, requirements, guidance (e.g., societies, publishers)
- Institutions/disciplinary services need to work together to help simplify workflows for authors (e.g., help desk, repositories)
- We, the community, need to find better ways to integrate data/software best practices earlier in the research process, embed in research workflows (e.g, platforms, notebooks)
- Researchers are inundated with guidance, we need to streamline information as much as possible in combination with the point above (e.g, checklists)
- We need to demonstrate to researchers the value of sharing data/software by leveraging metadata (e.g., filtering, indexing)

Thank you.

Chris Erdmann
Assistant Director, Data Leadership
American Geophysical Union
0000-0003-2554-180X
@libcce | cerdmann@agu.org