Data and Software: Important Building Blocks of Research

12 October 2021

Shelley Stall
Senior Director, Data Leadership
American Geophysical Union
0000-0003-2926-8353
@ShelleyStall | sstall@agu.org

AGU’s position statement on data affirms that

“Earth and space science data are a world heritage, and an essential part of the science ecosystem. All players in the science ecosystem—researchers, repositories, publishers, funders, institutions, etc.—should work to ensure that relevant scientific evidence is processed, shared, and used ethically, and is available, preserved, documented, and fairly credited.”

https://www.agu.org/Share-and-Advocate/Share/Policymakers/Position-Statements/Position_Data

AGU's position statement on data affirms that

"Earth and space science data are a **world heritage,** and an essential part of the science ecosystem. All players in the science ecosystem—researchers, repositories, publishers, funders, institutions, etc.—should work to **ensure that relevant scientific evidence is processed, shared, and used ethically, and is available, preserved, documented, and fairly credited.**"

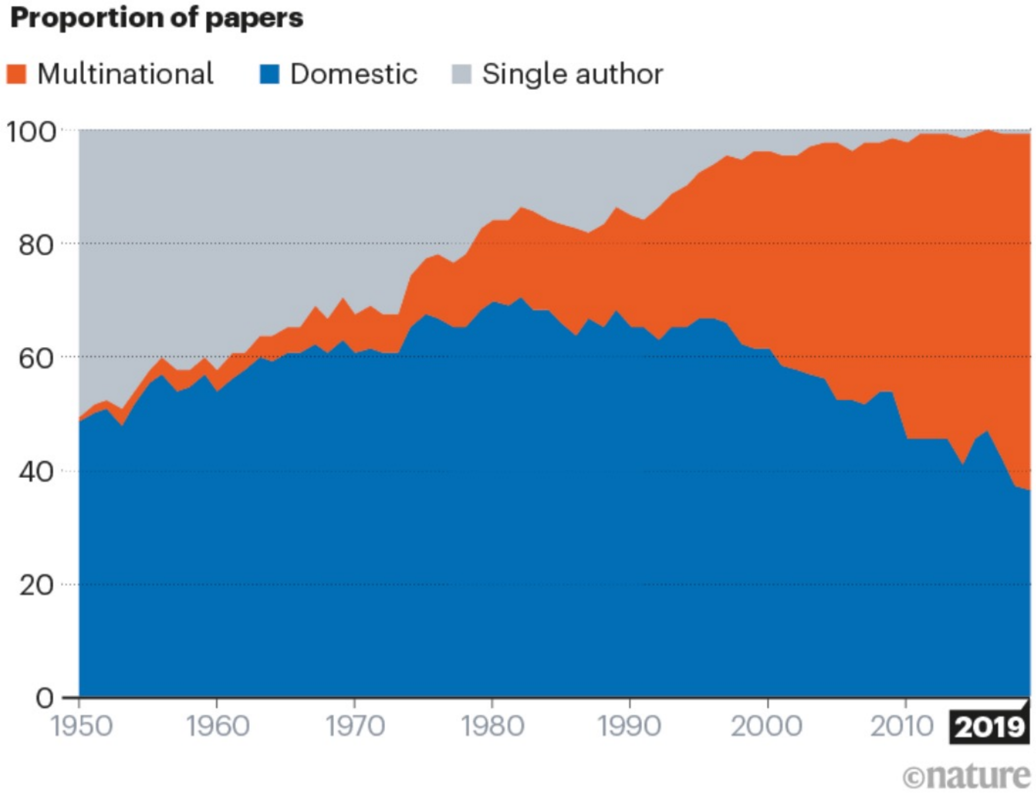
https://www.agu.org/Share-and-Advocate/Share/Policymakers/Position-Statements/Position_Data

150 years of *Nature*

A century and a half of research and discovery.

INTERNATIONAL COLLABORATIONS

Author lists on research publications show a shift towards multinational teams; fewer teams are composed entirely of researchers from one country.



Design: Jasiek Krzysztofiak; Data analysis: Digital Science Consultancy

Monastersky, R., & Van Noorden, R. (2019). 150 years of Nature: a data graphic charts our evolution. Nature, 575(7781), 22–23. https://doi.org/10.1038/d41586-019-03305-w

The Future of your Research

- Research Teams (not individuals)
- International Collaborations

- Robust **tools to discover** relevant research worldwide
- Good **documentation** to understand that research, data, and/or software

- Data that is **interoperable**, no matter which research team created it
- Software that is **accessible** and developed in current tools (e.g., Jupyter Notebooks)

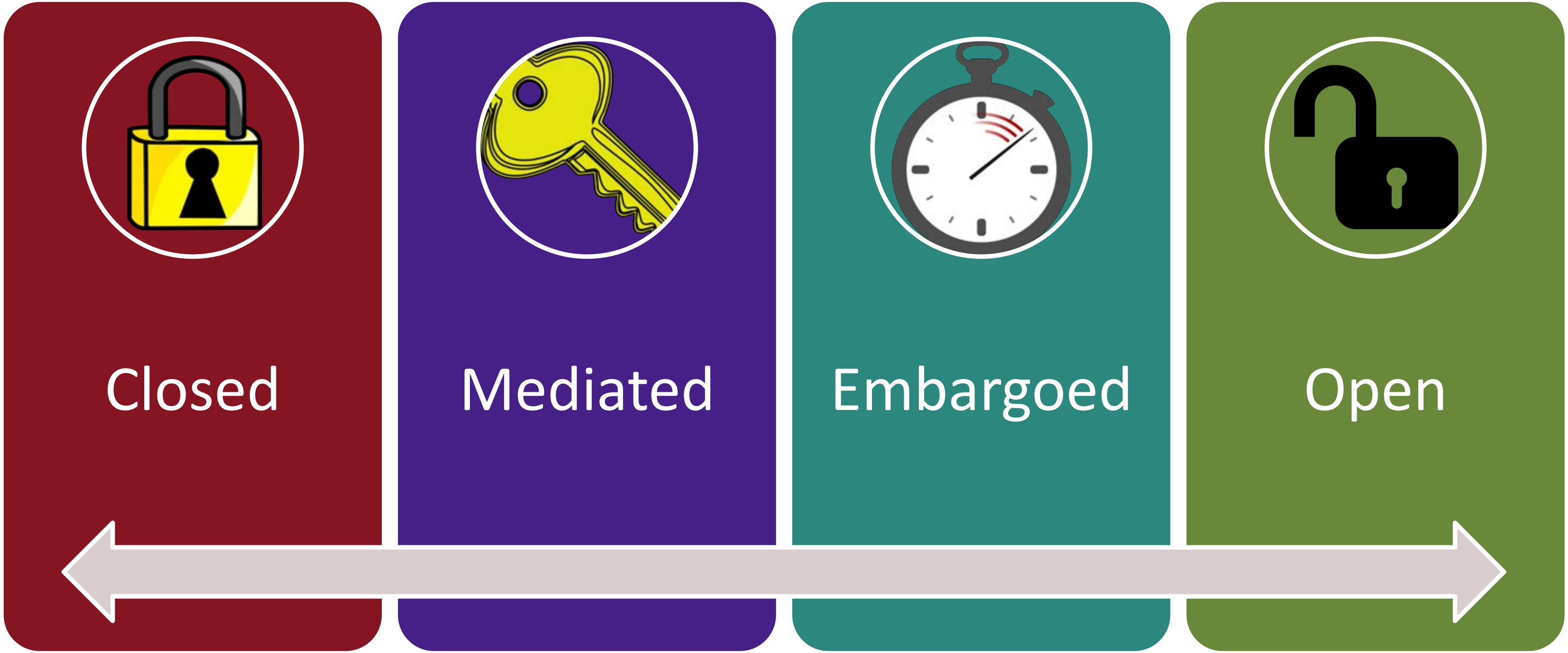
- Licenses that support **reuse**.

OPEN

Findable
Accessible
Interoperable
Reusable

FAIR

Closed vs Open – A continuum



Closed

Mediated

Embargoed

Open

FAIR Guiding Principles

FAIR is…

- **Findable**
- **Accessible**
- **Interoperable**
- **Reusable**

Article in Nature journal *Scientific Data*: Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).

FAIR Data Principles (applies to software and all digital objects)

- **Findable**
  - Assign persistent IDs (PIDs), provide rich metadata, register in a searchable resource, …

Repository

- **Accessible**
  - Retrievable by their ID using a standard protocol, metadata remain accessible even when data are no longer available…

Repository

- **Interoperable**
  - Use formal, broadly applicable languages, use standard vocabularies, qualified references…

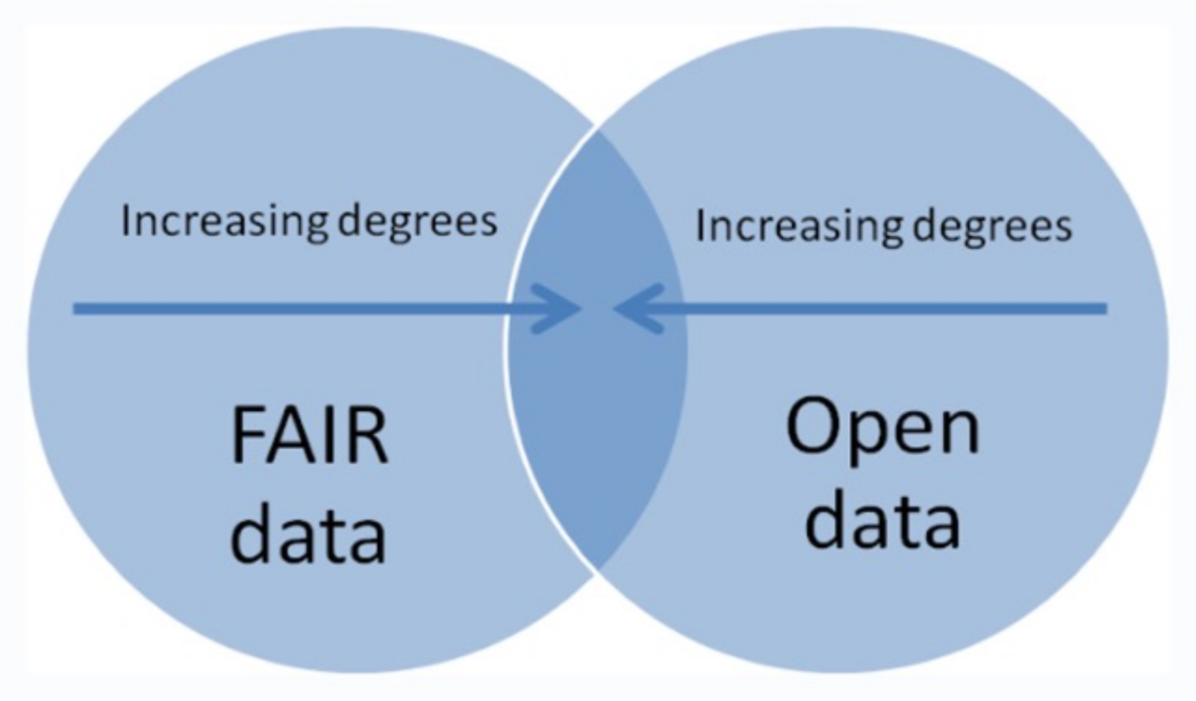
Community

- **Reusable**
  - Rich, accurate metadata, clear licenses, provenance, use of community standards…

Community / Repository

Article in Nature journal *Scientific Data*: Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).

Is FAIR Open? In short, "It depends."



Data can be FAIR or Open, both or neither.

The greatest potential for **reuse** comes when data are **both FAIR and Open.**

Higman, Rosie, Daniel Bangert, and Sarah Jones. 2019. "Three Camps, One Destination: The Intersections of Research Data Management, FAIR and Open". *Insights* 32 (1): 18. DOI: http://doi.org/10.1629/uksg.468

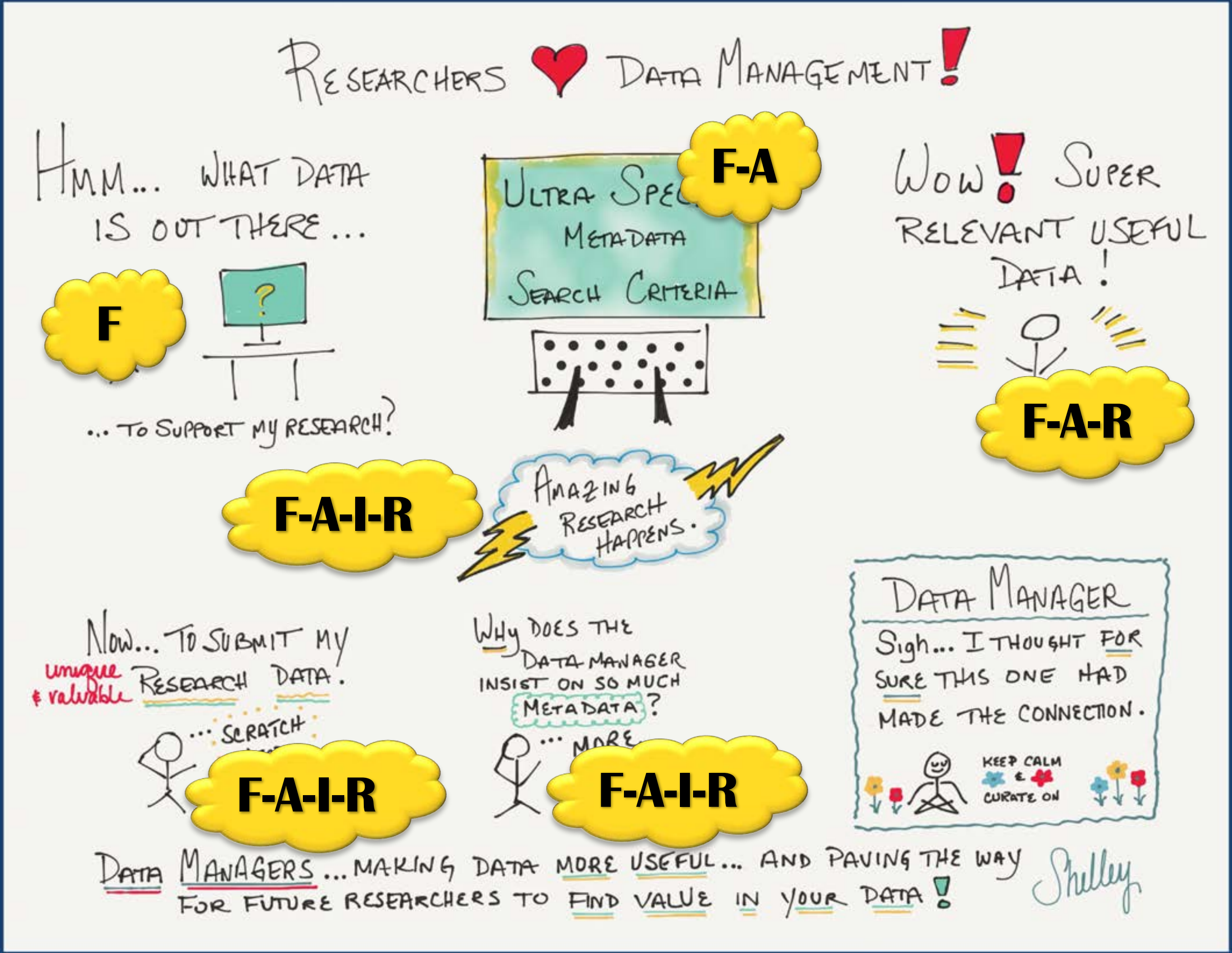
Data should be as open as possible, as closed as necessary.

What is the Continuum for FAIR? Two examples…

Persistent Identifier	◒ No unique ID	🟢 Globally resolvable persistent identifier
Robust Metadata	◒ No metadata	🟢 Metadata that aligns to community recommended standards/vocabulary/guidelines; Machine Readable

Why do we Care about FAIR?

ADVANCING EARTH AND SPACE SCIENCE


Researchers ❤ Data Management!
Hmm... what data is out there...
F
?
... to support my research?
Ultra Spec
F-A
Metadata
Search Criteria
Wow! Super relevant useful data!
F-A-R
F-A-I-R
Amazing research happens.
Now... to submit my unique & valuable research data.
... scratch
F-A-I-R
Why does the data manager insist on so much metadata?
... more
F-A-I-R
Data Manager
Sigh... I thought for sure this one had made the connection.
Keep calm & curate on
Data Managers... making data more useful... and paving the way for future researchers to find value in your data!
Shelley

Reprise: Annotated with FAIR elements

The Future of your Research

- Research Teams (not individuals)
- International Collaborations

- Robust **tools to discover** relevant research worldwide F-A
- Good **documentation** to understand that research, data, and/or software F-A-R

- Data that is **interoperable**, no matter which research team created it
- Software that is **accessible** and developed in current tools (e.g., Jupyter Notebooks)

F-A-I-R

- Licenses that support **reuse**. R

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

OPEN SCIENCE BY DESIGN

Realizing a Vision for 21st Century Research

National Academies of Sciences, Engineering, and Medicine. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25116.

Open Science by Design is aimed at overcoming barriers and **moving toward open science as the default approach across the research enterprise**. This report explores specific examples of open science and discusses a range of challenges, focusing on stakeholder perspectives. It is meant to provide guidance to the research enterprise and its stakeholders as they build strategies for achieving open science and take the next steps.

Report released July 2018

AGU 100
ADVANCING EARTH AND SPACE SCIENCE

NASEM Definition of Open Science

Open science aims to ensure the open availability and usability:

- scholarly publications
- the data that result from scholarly research
- and the methodologies, including code or algorithms, that were used to generate those data.

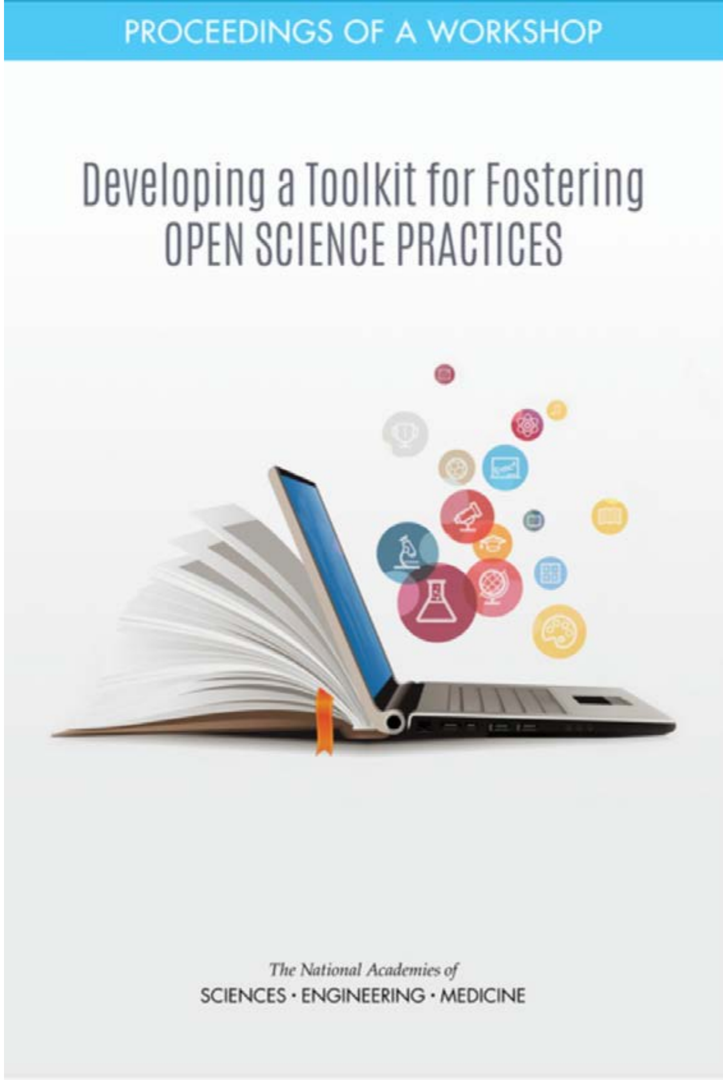
5 Findings and Recommendations

1. Building a Supportive Culture - internationally
2. Training for Open Science by Design
3. Ensuring long term preservation and stewardship - Developing and sustaining the infrastructure
4. Facilitating data discovery, and reuse – FAIR Data Principles
5. Developing new approaches to fostering Open Science by Design – with public and private funder support

Developing a Toolkit for Fostering Open Science Practices

Released 30 Sept 2021 from The National Academies Roundtable on Aligning Incentives for Open Science

National Academies of Sciences, Engineering, and Medicine. 2021. *Developing a Toolkit for Fostering Open Science Practices: Proceedings of a Workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26308.



PROCEEDINGS OF A WORKSHOP

Developing a Toolkit for Fostering
OPEN SCIENCE PRACTICES

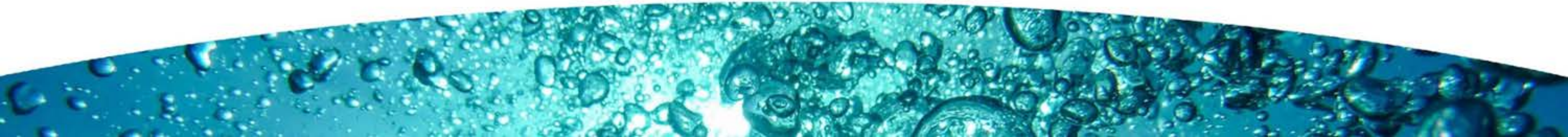
The National Academies of
SCIENCES · ENGINEERING · MEDICINE

Open Science – The Key Pillars

- Open scientific knowledge
  - **Scientific publications**
  - **Open research data**
  - **Open source software and source code**
  - **Open hardware**
- Open science infrastructures
- Science communication
- Open engagement of societal actors
- Open dialogue with other knowledge systems.

Source: Final report on the draft text of the UNESCO Recommendation on Open Science; final adoption expected November 2021

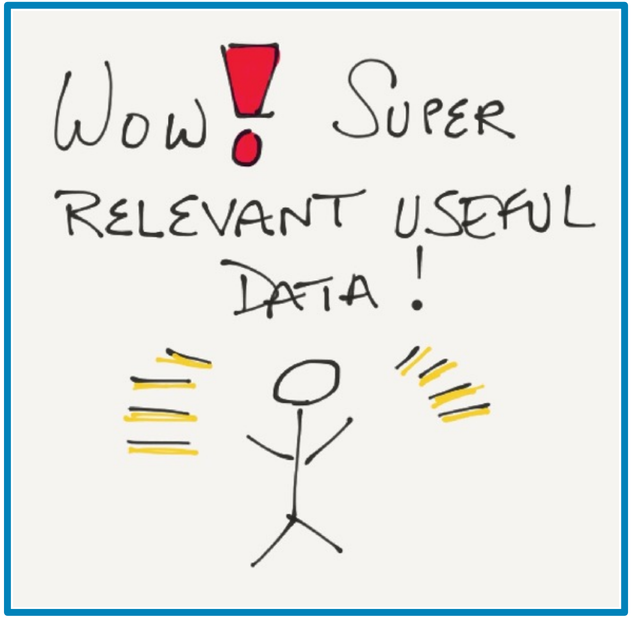
https://en.unesco.org/science-sustainable-future/open-science



From the very beginning of the research process,

the researcher **both contributes** to open science and

takes advantage of the open science practices of other members of the research community.



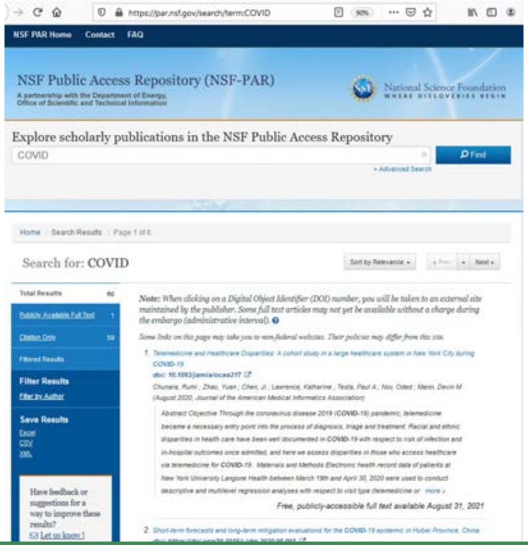
For Publications

NSF Public Access Repository (PAR) 1.0 Current System Functions and Characteristics

PAR 1.0 focused on Peer-Reviewed Articles

- Enables researchers to enter metadata for peer-reviewed articles or auto-populate by means of Digital Object Identifiers (DOI)
- Metadata recorded in PAR is also transmitted and synchronized with Award Search database
- The metadata for articles can be searched and displayed
- Researchers may deposit (or retrieve) public access articles as PDF/A files or point to download locations through DOI
- Note that PAR is not a single system, but rather an infrastructure of various interacting software systems in several different parts of the NSF infrastructure, also making use of modified modules from the DOE Office of Scientific and Technical Information (OSTI) system

URL: http://par.nsf.gov



NSF Public Access Repository (NSF-PAR)
A partnership with the Department of Energy, Office of Scientific and Technical Information

Explore scholarly publications in the NSF Public Access Repository

COVID

Search for: COVID

Total Results	60
Publicly Available Full Text	1
Citation Only	59
Filtered Results	

Note: *When clicking on a Digital Object Identifier (DOI) number, you will be taken to an external site maintained by the publisher. Some full text articles may not yet be available without a charge during the embargo (administrative interval).*

Some links on this page may take you to non-federal websites. Their policies may differ from this site.

1. Telemedicine and Healthcare Disparities: A cohort study in a large healthcare system in New York City during COVID-19
   doi: 10.1093/jamia/ocaa217
   Chunara, Rumi ; Zhao, Yuan ; Chen, Ji ; Lawrence, Katharine ; Testa, Paul A ; Nov, Oded ; Mann, Devin M
   (August 2020, Journal of the American Medical Informatics Association)

   Abstract Objective Through the coronavirus disease 2019 (COVID-19) pandemic, telemedicine became a necessary entry point into the process of diagnosis, triage and treatment. Racial and ethnic disparities in health care have been well documented in COVID-19 with respect to risk of infection and in-hospital outcomes once admitted, and here we assess disparities in those who access healthcare via telemedicine for COVID-19 . Materials and Methods Electronic health record data of patients at New York University Langone Health between March 19th and April 30, 2020 were used to conduct descriptive and multilevel regression analyses with respect to visit type (telemedicine or more »

   Free, publicly-accessible full text available August 31, 2021

2. Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province, China

Have feedback or suggestions for a way to improve these results?
Let us know!

Slide Credit: Martin Halbert, NSF Senior Advisor for Public Access, 19 November 2020

NSF PAR 2.0 Development Planning

- Primary goal for the second version of PAR is to accommodate submissions of research data sets
- Will not require that data sets be deposited in PAR 2.0 but instead will expect data sets to be deposited in repositories which demonstrate FAIR best practices such as maintaining sustainable long-term access through the assignment of a persistent identifier (PID) with quality descriptive metadata
- Now planning the upgrades to the NSF PAR infrastructure to address the additional needed features
- Aim to accomplish major development efforts in calendar 2021

Slide Credit: Martin Halbert, NSF Senior Advisor for Public Access, 19 November 2020

Slide Credit: Susan K. Gregurick, Ph.D., Associate Director for Data Science
National Institutes of Health, 19 November 2020

Final NIH Policy for Data Management and Sharing (NOT-OD-21-013)

Release Date: **October 29, 2020 |** Effective Date: **January 25, 2023**

NIH requires researchers to prospectively plan for how scientific data will be preserved and shared through submission of a Data Management and Sharing Plan

Submission of a Data Management and Sharing Plan outlining how scientific data and any accompanying metadata will be managed and shared, taking into account any potential restrictions or limitations.

Plan is part of the budget Justification section of the application for extramural awards and as part of the technical evaluation for contracts

The DMS Policy applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data. This includes research funded or conducted by extramural grants, contracts, Intramural Research Projects, or other funding agreements regardless of NIH funding level or funding mechanism. **The DMS Policy does not apply to research and other activities that do not generate scientific data, including training, infrastructure development, and non-research activities.**

Final NIH Policy for Data Management and Sharing

NOT-OD-21-015: Allowable Costs for Data Management and Sharing include curation, data management and sharing data through repositories

- Shared scientific data should be made accessible as soon as possible, and no later than the time of an associated publication, or the end of performance period, whichever comes first.

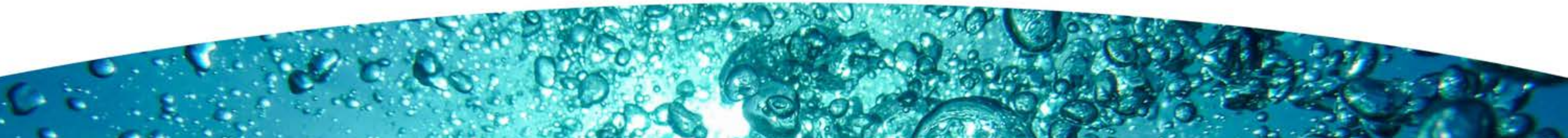
NOT-OD-21-016: Selecting a Repository include desirable characteristics for a data sharing repository

- **NIH strongly encourages the use of established repositories** to the extent possible for preserving and sharing scientific data.

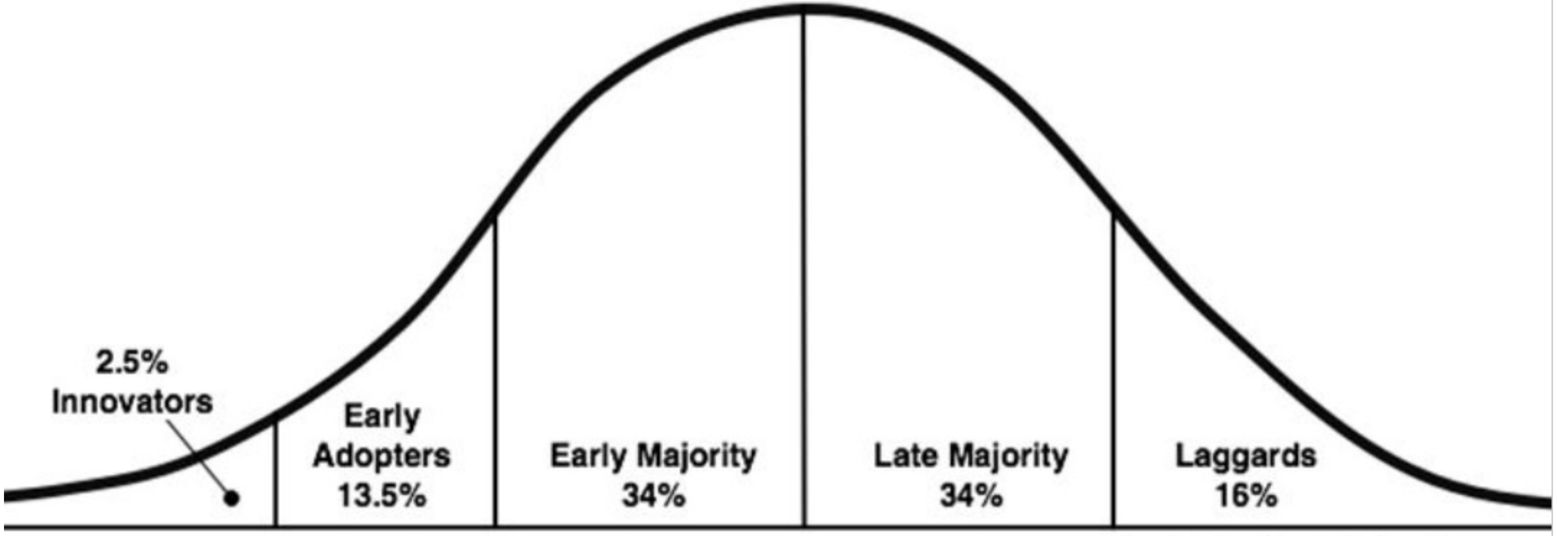
Slide Credit: Susan K. Gregurick, Ph.D., Associate Director for Data Science
National Institutes of Health, 19 November 2020

What is the Funder Focus?

1. Data Management Plans
   - PI required to consider **how** data will managed and preserved (i.e., repository selection)
2. Linking Data (and software) to a grant, publication, researcher, etc
   - Persistent Identifiers (**PIDs**) and linked infrastructure
3. Culture change
   - Evolving from recommendations to requirements



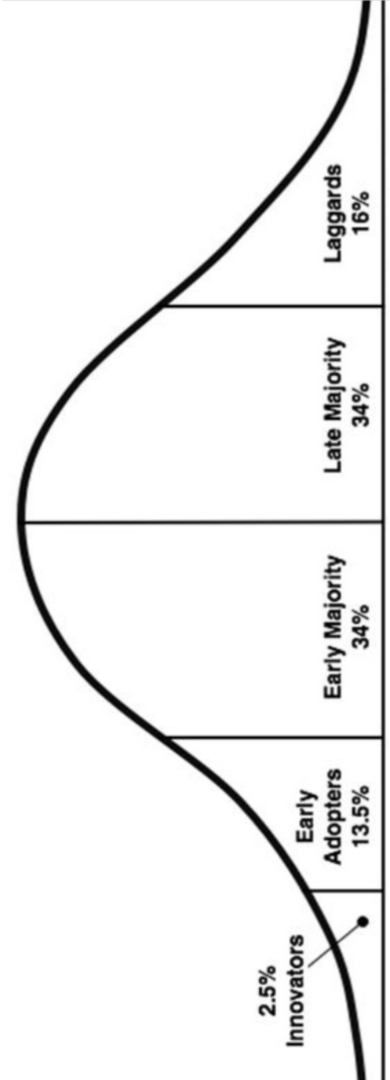
Changing a Research Culture



Source: Everett Rogers *Diffusion of innovations model*

Diffusion of Innovations; Rogers, 1963

Adapted from Brian Nosek (COS) by Marcia McNutt, President of the National Academy of Science



Source: Everett Rogers Diffusion of innovations model

Make it required – with policies

Make it rewarding – with incentives

Make it accepted – by building communities

Community of Practice: Promote data sharing, metadata standards, criteria for interoperability

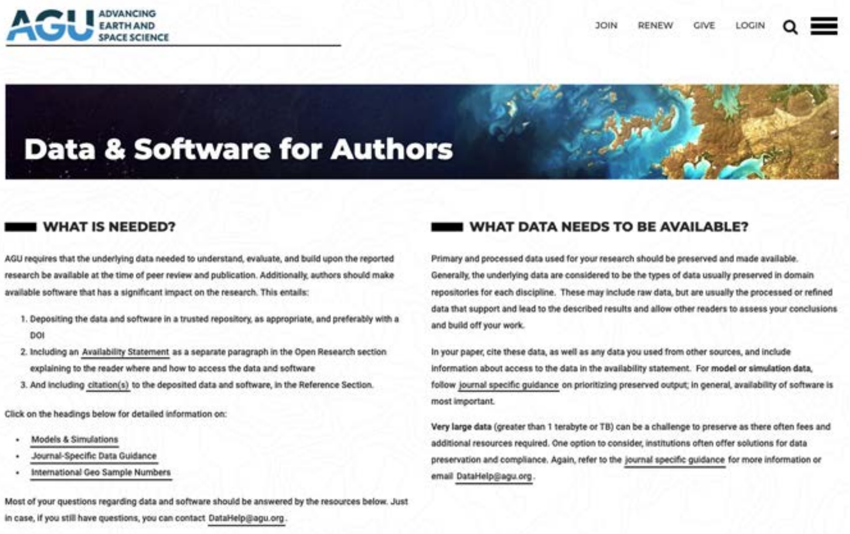
Make it easy – with user interface

Make it possible – with infrastructure

AGU Data & Software Sharing Guidance

What is covered:

- What data needs to be available?
- Repository Selection
- Availability Statement
- Data & Software Citation
- Physical Samples



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.

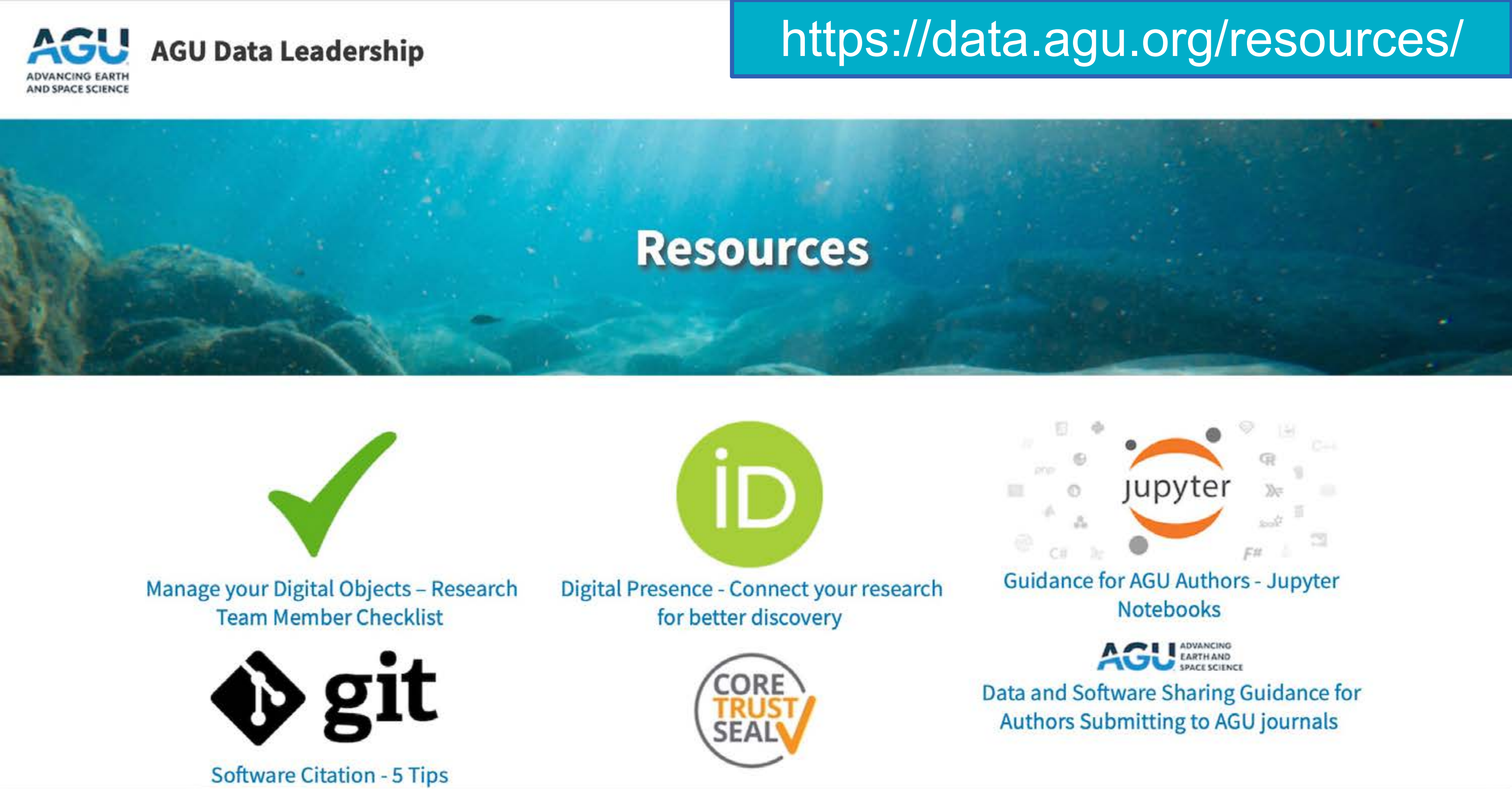
WHAT DATA NEEDS TO BE AVAILABLE?

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 https://data.agu.org/resources/



AGU Data Leadership

Resources

Manage your Digital Objects – Research Team Member Checklist

Digital Presence - Connect your research for better discovery

Guidance for AGU Authors - Jupyter Notebooks

Software Citation - 5 Tips

Data and Software Sharing Guidance for Authors Submitting to AGU journals

Thank you

Shelley Stall
AGU Sr. Director, Data Leadership
sstall@agu.org
@ShelleyStall
https://orcid.org/0000-0003-2926-8353

Findable Accessible Interoperable Reusable
Make Data FAIR
AGU100 ADVANCING EARTH AND SPACE SCIENCE

This work is generously funded through grants from:

FAPESP FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO
JST
PARSEC
BELMONT FORUM
NSF
UNESCO
AGENCE NATIONALE DE LA RECHERCHE ANR
AGU ADVANCING EARTH AND SPACE SCIENCE