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.

Proportion of papers

- Multinational
- Domestic
- Single author

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

FAIR

Findable
Accessible
Interoperable
Reusable
Closed vs Open – A continuum

Closed  Mediated  Embargoed  Open
FAIR Guiding Principles

FAIR is...
Findable
Accessible
Interoperable
Reusable

Article in Nature journal *Scientific Data*:
FAIR Data Principles (applies to software and all digital objects)

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

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

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

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

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.

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?
Researchers ❤ Data Management!

HMM... WHAT DATA IS OUT THERE...

... TO SUPPORT MY RESEARCH?

F-A-I-R

ULTRA SPECIFIC METADATA SEARCH CRITERIA

F-A

NOW! SUPER RELEVANT USEFUL DATA!

F-A-R

F-A-I-R

AMAZING RESEARCH HAPPENS.

F-A-I-R

Now... To submit my unique & valuable research data.

Why does the data manager insist on so much metadata?

F-A-I-R

Data manager... making data more useful... and paving the way for future researchers to find value in your data!

Shelley
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 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
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

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

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.
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
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
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.
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

Diffusion of Innovations; Rogers, 1963
Adapted from Brian Nosek (COS) by Marcia McNutt, President of the National Academy of Science

- Make it required – with policies
- Make it rewarding – with incentives
- Make it accepted – by building communities
- Make it easy – with user interface
- Make it possible – with infrastructure

Community of Practice: Promote data sharing, metadata standards, criteria for interoperability
AGU Data & Software Sharing Guidance

What is covered:

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

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

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

This work is generously funded through grants from: