SC23

Denver, CO | i am hpc.

FAIRIST of them all: Meeting researchers where they are with just-in-time, FAIR implementation advice

Christine Kirkpatrick
Division Director, Research Data Services
San Diego Supercomputer Center,
UC San Diego

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

U.PORTO FC FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

GO FAIR US Advancing FAIR in the US

SDSC SAN DIEGO SUPERCOMPUTER CENTER

This work is supported through the NSF awards #2226453, 2138811, 1916481.
Image source: Dynabench's Adversarial Nibbler

5/3/23

The Promised Land

- Scientific impact with maximum ease
- Optimized resource utilization
- Effective self-service, self-healing resources
- Reusable, reproducible, open science principles and values

Strange and barely relevant images from: Dynabench's Adversarial Nibbler

Challenges

- Proverbial 80% of time with data is spent finding and cleaning it
- Increased requirements from funding agencies, publishers, and institutions
- Rapid innovation alongside difficulty identifying useful or applicable practices
- Even when one defines a great Data Management and Sharing Plan, implementing and sustaining it
- Ethics and/or things we should do but aren’t required to do (yet)

How Data Scientists Spend Their Time

Building training sets	3%
Cleaning and organizing data	***60%***
Collecting data sets	***19%***
Mining data for patterns	9%
Refining algorithms	4%
Other	5%

‘Data Scientists Time’ Source: Data Science Report 2016, CrowdFlower, 2016

New Disruptors

Everything AI

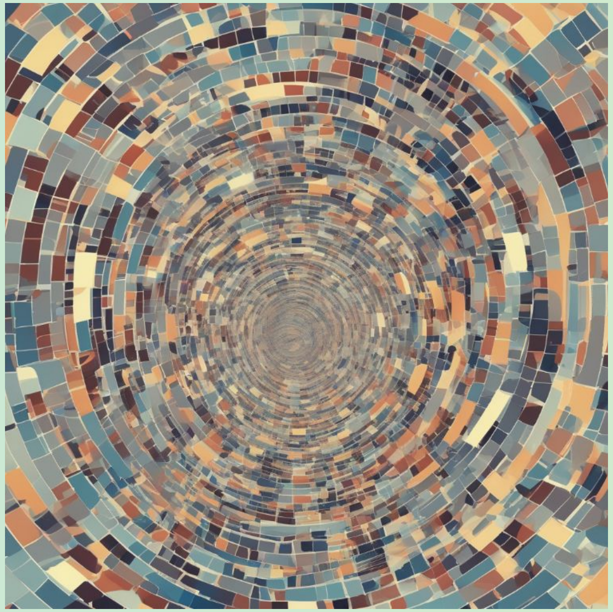
- Untangling the hype
- Building foundation models
- Generative AI and resources
- AI reproducibility
- Student demand and focus

Images source: Dynabench's Adversarial Nibbler
Prompt: disruptive AI technology and student demand

Everything Old is New Again

Data-driven research (even in an AI context) still requires:

- Well-annotated data
- Reliable tools
- Accessible, extensible infrastructure
- Benchmarking practices
  - to innovate your own infrastructure
  - architect purpose-built systems
- Training and education
  - just-in-time information
  - understanding of ethical implications for new technologies
  - support to choose relevant new tools and methods and to take advantage of new knowledge
- The culture and resources to support all of the above



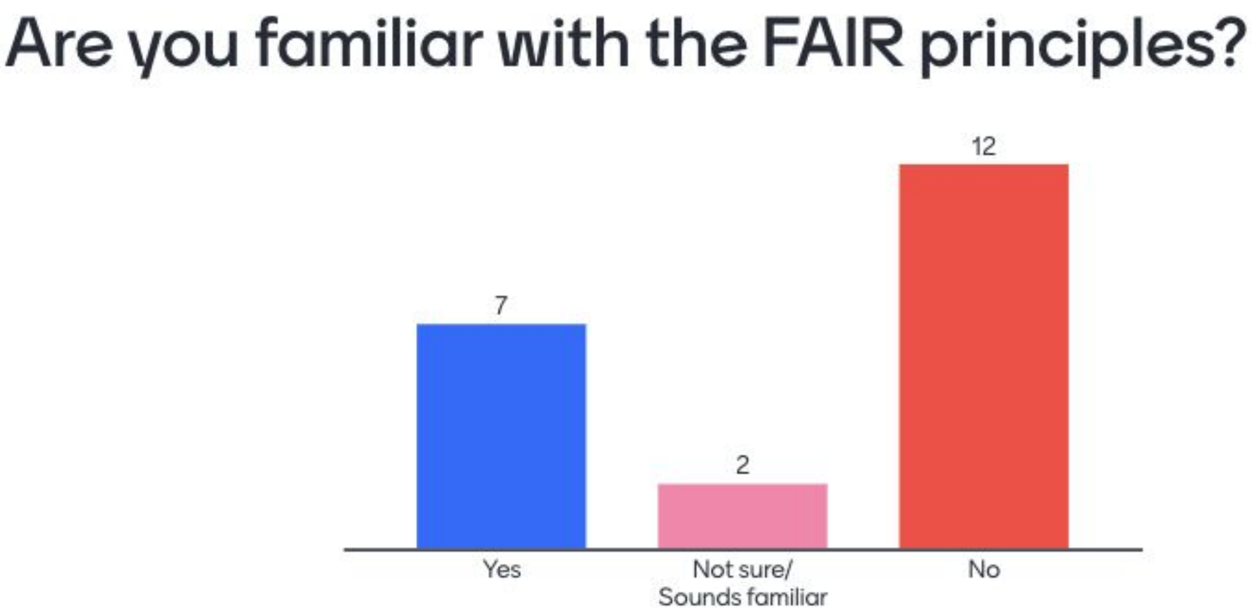
Images source: Dynabench's Adversarial Nibbler
Prompt: data-driven research requires well described, machine actionable data

Overview

1. Landscape Context & Primer
    a. FAIR Principles
    b. FAIR Digital Objects
    c. Open Science
    d. AI Readiness
2. Putting Everything into Practice
    a. SDSC Data Journey
    b. FAIRIST
    c. FARR
    d. Future Work

FAIR Principles

- 15 Principles, not a specification
- SC ‘22 poll showed HPC data divide →
- Not just for data!
  Come to the SDSC booth Wednesday at 2:30 for Sean Wilkinson’s talk on FAIR Workflows
- Continuous goal, not a destination
- Spectrum - good/better/best
- Machine actionability
- Required by funders at proposal stage



Findable

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. metadata specify the data identifier.
F4. (meta)data are registered or indexed in a searchable resource.

Interoperable

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

Accessible

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

Reusable

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

Wilkinson, 2016. https://doi.org/10.1038/sdata.2016.18

How to FAIR 101

F Assigning unique identifiers to your data (PID, DOI) include in the metadata record

F Metadata should be machine actionable

F Registering your data or depositing in data repositories

A Provide an API or web-based mechanism for querying at least the metadata

I Use standard vocabularies, taxonomies, or ontologies that are documented on fairsharing.org or BioPortal

I+ Recording provenance in accompanying metadata

I+ Documenting software needed to use the data, including providing access to the software on GitHub, etc.

R Include clear information on the data license (choose one at Creative Commons)

R Include provenance in the metadata, and how to cite the resource

R Follow documented standards, e.g., diseases map to ICD-11

R+ Provide a (Jupyter/R) notebook

+ contributes to reproducibility

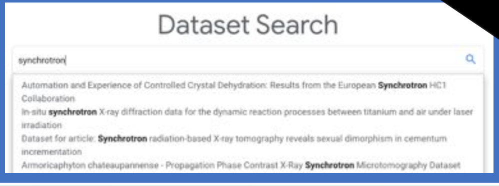
Baked in and/or Layered Metadata

Semantic Web

- 2006, Tim Berners-Lee
- RDF / linked data

Schema.org

- JSON-LD
- Google Datasets


Dataset Search

synchrotron

Automation and Experience of Controlled Crystal Dehydration: Results from the European **Synchrotron** HC1 Collaboration

In-situ **synchrotron** X-ray diffraction data for the dynamic reaction processes between titanium and air under laser irradiation

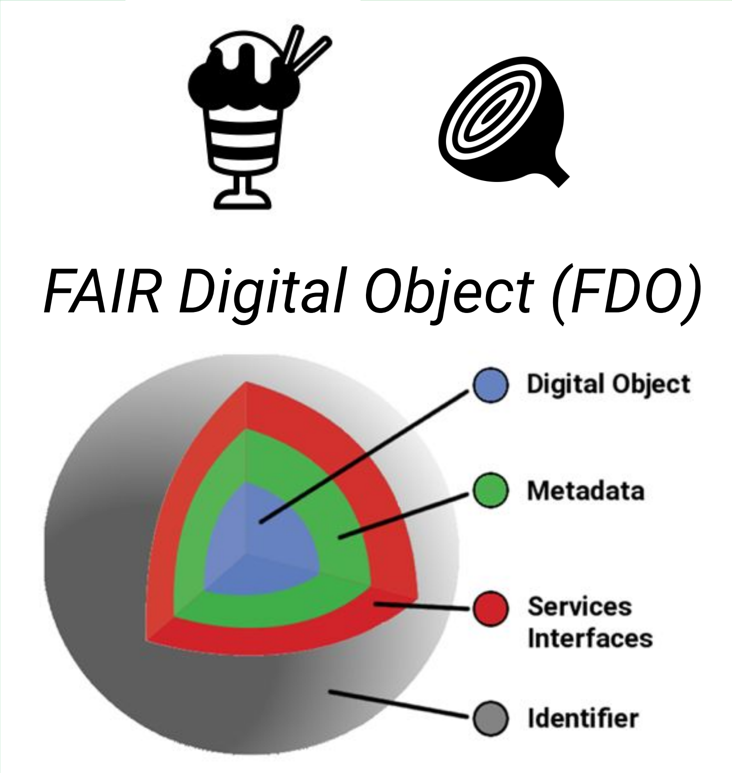
Dataset for article: **Synchrotron** radiation-based X-ray tomography reveals sexual dimorphism in cementum incrementation

Armoricaphyton chateaupannense - Propagation Phase Contrast X-Ray **Synchrotron** Microtomography Dataset

Community Extensions

Bioschemas

ESIPFed / science-on-schema.org


FAIR Digital Object (FDO)

FAIR DIGITAL OBJECTS FORUM

http://fairdo.org

Digital Object graphic: (Schwardmann 2020)

THE WHITE HOUSE
WASHINGTON

YEAR OF OPEN SCIENCE

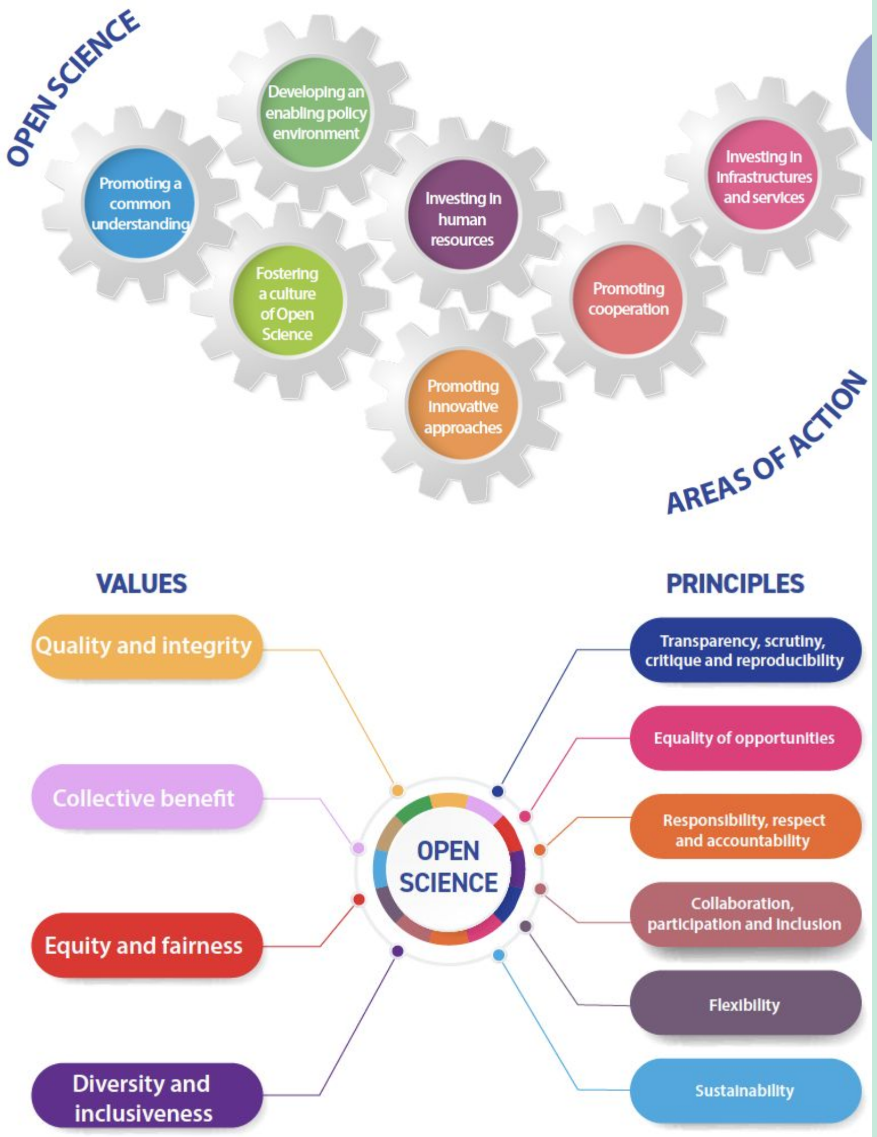
✦ NASA ✦ NSF ✦ NOAA ✦
✦ DOE ✦ GSA ✦ NEH ✦ NIH ✦
✦ NIST ✦ USDA ✦ USGS ✦

White House definition:
Open and equitable research

- Special funding calls from NASA TOPS
- Updating agency public access plans per the Nelson memo
- Roll out of NIH Data Management & *Sharing* Requirements
- NSF GEO OSE program

Nelson Memo: Aug. 2022 from OSTP calls for agencies **update public access policies** & implement plans **no later than 2025**, to **end data embargoes, data available free and immediately** by default

UNESCO’s Recommendation on Open Science



OPEN SCIENCE

Promoting a common understanding

Developing an enabling policy environment

Fostering a culture of Open Science

Investing in human resources

Promoting innovative approaches

Promoting cooperation

Investing in infrastructures and services

AREAS OF ACTION

VALUES

Quality and integrity

Collective benefit

Equity and fairness

Diversity and inclusiveness

OPEN SCIENCE

PRINCIPLES

Transparency, scrutiny, critique and reproducibility

Equality of opportunities

Responsibility, respect and accountability

Collaboration, participation and inclusion

Flexibility

Sustainability

Watch the recording from the **National Science Data Fabric Distinguished Lecture series:**

Dr. Ana Peršić

Science Policy and Partnerships Section, Division of Science Policy and Capacity Building, UNESCO

‘The pathway to implementing the UNESCO Recommendation on Open Science’

https://nationalsciencedatafabric.org → Seminars

http://on.unesco.org/OpenScience

AI Readiness

- Current literature focuses on AI readiness for organizations
- Everything we learned from making data SQL-ready
- Cleaned up data
  - True/False → 1/0
  - Punctuation removed (esp. punctuation that breaks code)
- Well described data
  - Documented, controlled vocabularies
  - Taxonomies available for ‘super’ categories
    - Residence, apartment = domicile
- Technology aware
  - RDF/linked data for use in knowledge graphs

The World According to Christine

This semi-relevant graphic brought to you by Dynabench’s adversarial nibbler

Putting Everything into Practice

San Diego Supercomputer Center

Founded in 1985

19K sq ft data center, 3.5 MW

Flagship Systems

- Expanse - 5 petaflops
- Voyager - optimized for Deep Learning

Research Data Services

- Everything around HPC
- Networking, platforms, storage, cloud, project support
- Research data management + research computing
- Innovative data-driven initiatives

GO FAIR US

Advancing FAIR in the US

WEST BIG DATA INNOVATION HUB

EARTHCUBE

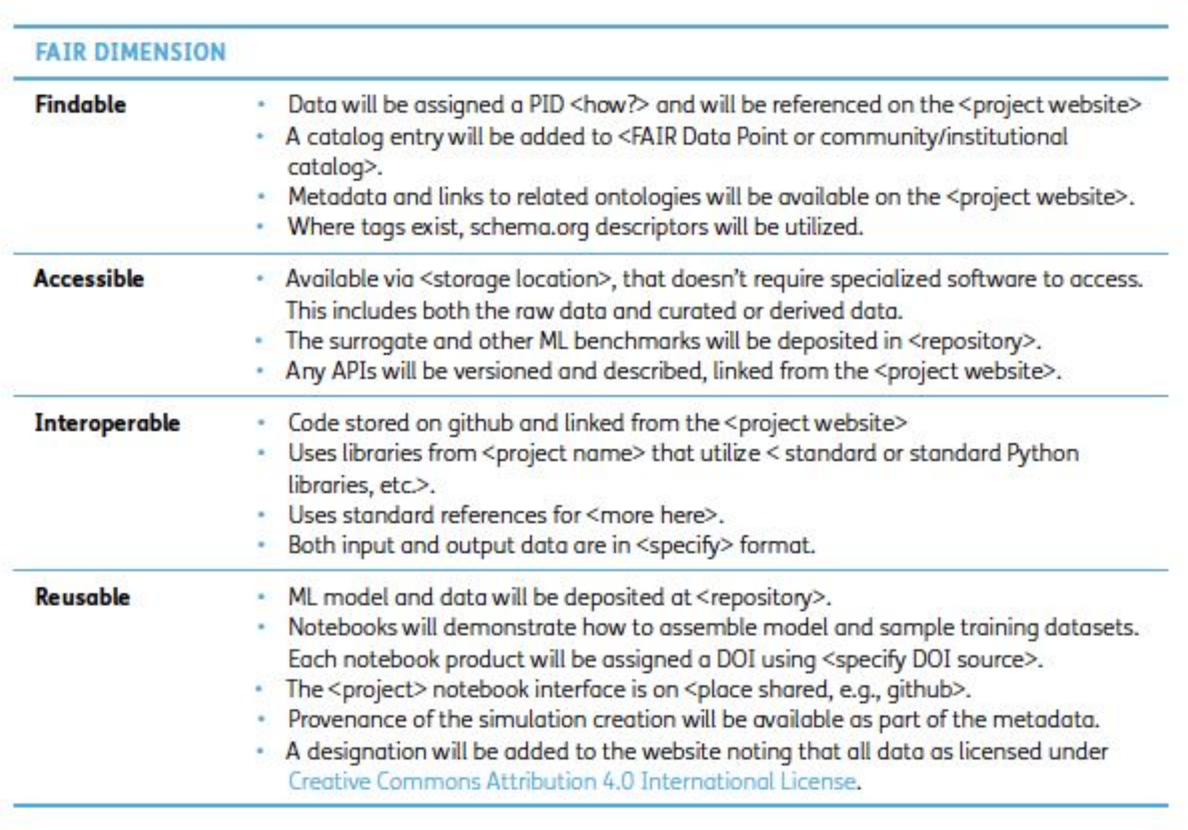
TRANSBOUNDARY GROUNDWATER RESILIENCE

Funded by the National Science Foundation

WATER SOCIAL DATA SYSTEMS

SDSC/RDS: Our Journey to Leading in Data

- Researcher: I need storage and a VM
  Translation: create a dataset
- GO FAIR training
  - First data stewardship week in 2018
  - Train-the-trainer event in February 2020
- CODATA, Secretary General (2021-2025)
- GO FAIR US, Head
- National Academies Committees
  - Board on Research Data and Information (BRDI)
  - US National Committee on CODATA
- Research Data Alliance
  - Organizational Assembly
  - Technical Advisory Board (2018-2021)
- Best chance for a conversation about FAIR practices is during the proposal phase



FAIR DIMENSION	
Findable	• Data will be assigned a PID <how?> and will be referenced on the <project website> • A catalog entry will be added to <FAIR Data Point or community/institutional catalog>. • Metadata and links to related ontologies will be available on the <project website>. • Where tags exist, schema.org descriptors will be utilized.
Accessible	• Available via <storage location>, that doesn't require specialized software to access. This includes both the raw data and curated or derived data. • The surrogate and other ML benchmarks will be deposited in <repository>. • Any APIs will be versioned and described, linked from the <project website>.
Interoperable	• Code stored on github and linked from the <project website> • Uses libraries from <project name> that utilize < standard or standard Python libraries, etc.>. • Uses standard references for <more here>. • Both input and output data are in <specify> format.
Reusable	• ML model and data will be deposited at <repository>. • Notebooks will demonstrate how to assemble model and sample training datasets. Each notebook product will be assigned a DOI using <specify DOI source>. • The <project> notebook interface is on <place shared, e.g., github>. • Provenance of the simulation creation will be available as part of the metadata. • A designation will be added to the website noting that all data as licensed under Creative Commons Attribution 4.0 International License.

Example table supplied to researchers for their DMP

Kirkpatrick CR, Coakley K, Christopher J, Dutra I. Engaging with Researchers and Raising Awareness of FAIR and Open Science through the FAIR+ Implementation Survey Tool (FAIRIST). Data Science Journal. 2023; 22:32. Available from: https://datascience.codata.org/articles/10.5334/dsj-2023-032

Turning Point: NIH Data Sharing and Management Plan

Researchers need to know:

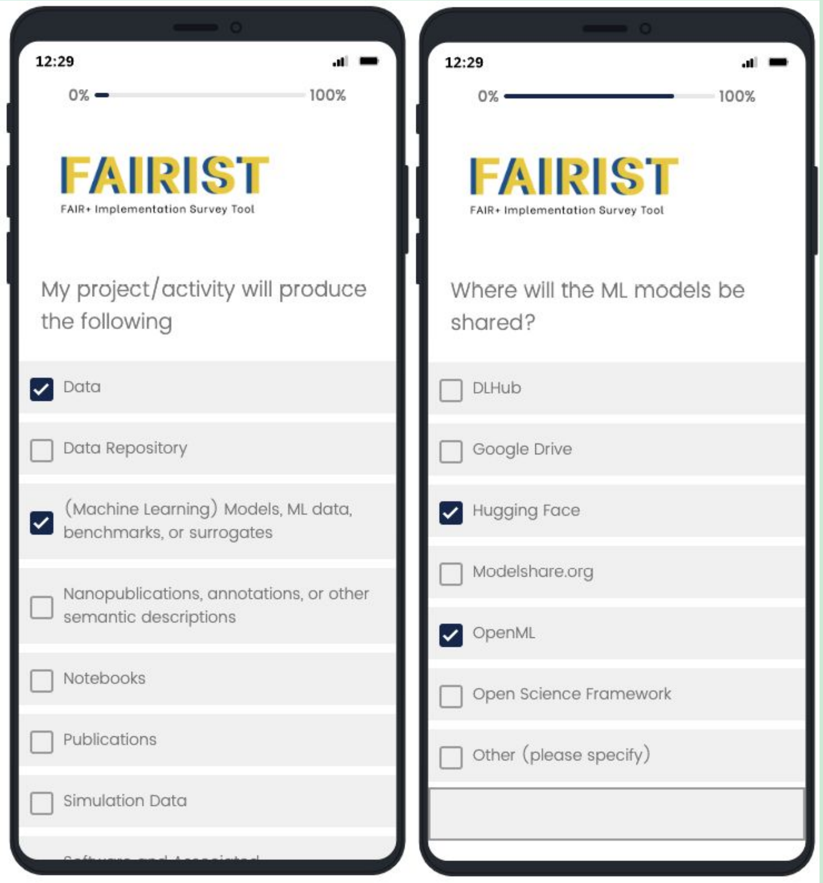
- the metadata they plan to use
- what standards will be used for metadata and data
- the repository they will deposit data in
- a plan for unique identifiers

Key questions:

- Summarize the **types** and estimated amount of scientific data expected to be generated in the project
- **Briefly list the metadata**, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) **that will be made accessible to facilitate interpretation of the scientific data**.
- State **what common data standards will be applied to the scientific data** and associated metadata to enable interoperability of datasets and resources, and **provide the name(s) of the data standards** that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project.
- **Provide the name of the repository**(ies) where scientific data and metadata arising from the project will be archived
- Describe **how the scientific data will be findable and identifiable**, i.e., via a persistent unique identifier or other standard indexing tools.
- Describe and **justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse** of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing.
- Describe how compliance with **this Plan will be monitored and managed**, frequency of oversight, and by whom at your institution

FAIRIST: FAIR+ Implementation Survey Tool

- FAIR + reproducibility
  - AI practices
- Convert what I know into rules
- Reduce FAIR implementation into if/then
- Augment tool as practices are developed
- Use Turbotax like interface with almost no fill in the blank
- Provide links and just-in-time information relevant to the project
- Proof of concept
  → Amend other tools
- Try it out at fairist.sdsc.edu
- Feedback tinyurl.com/fairist



12:29

0% 100%

FAIRIST

FAIR+ Implementation Survey Tool

My project/activity will produce the following

- ☑ Data
- □ Data Repository
- ☑ (Machine Learning) Models, ML data, benchmarks, or surrogates
- □ Nanopublications, annotations, or other semantic descriptions
- □ Notebooks
- □ Publications
- □ Simulation Data

12:29

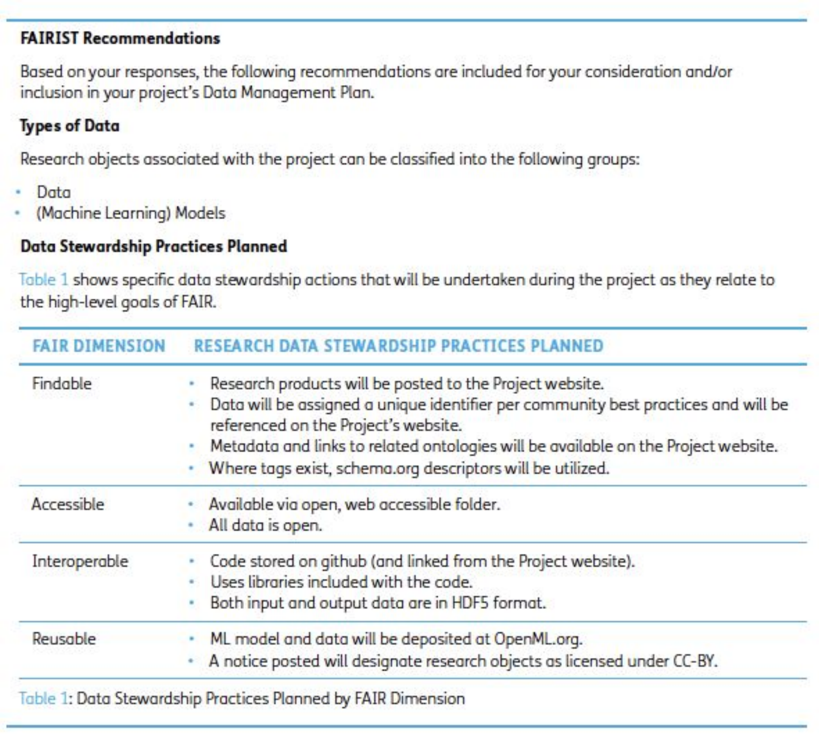
0% 100%

FAIRIST

FAIR+ Implementation Survey Tool

Where will the ML models be shared?

- □ DLHub
- □ Google Drive
- ☑ Hugging Face
- □ Modelshare.org
- ☑ OpenML
- □ Open Science Framework
- □ Other (please specify)



FAIRIST Recommendations

Based on your responses, the following recommendations are included for your consideration and/or inclusion in your project's Data Management Plan.

Types of Data

Research objects associated with the project can be classified into the following groups:

- Data
- (Machine Learning) Models

Data Stewardship Practices Planned

Table 1 shows specific data stewardship actions that will be undertaken during the project as they relate to the high-level goals of FAIR.

FAIR DIMENSION	RESEARCH DATA STEWARDSHIP PRACTICES PLANNED
Findable	• Research products will be posted to the Project website. • Data will be assigned a unique identifier per community best practices and will be referenced on the Project's website. • Metadata and links to related ontologies will be available on the Project website. • Where tags exist, schema.org descriptors will be utilized.
Accessible	• Available via open, web accessible folder. • All data is open.
Interoperable	• Code stored on github (and linked from the Project website). • Uses libraries included with the code. • Both input and output data are in HDF5 format.
Reusable	• ML model and data will be deposited at OpenML.org. • A notice posted will designate research objects as licensed under CC-BY.

Table 1: Data Stewardship Practices Planned by FAIR Dimension

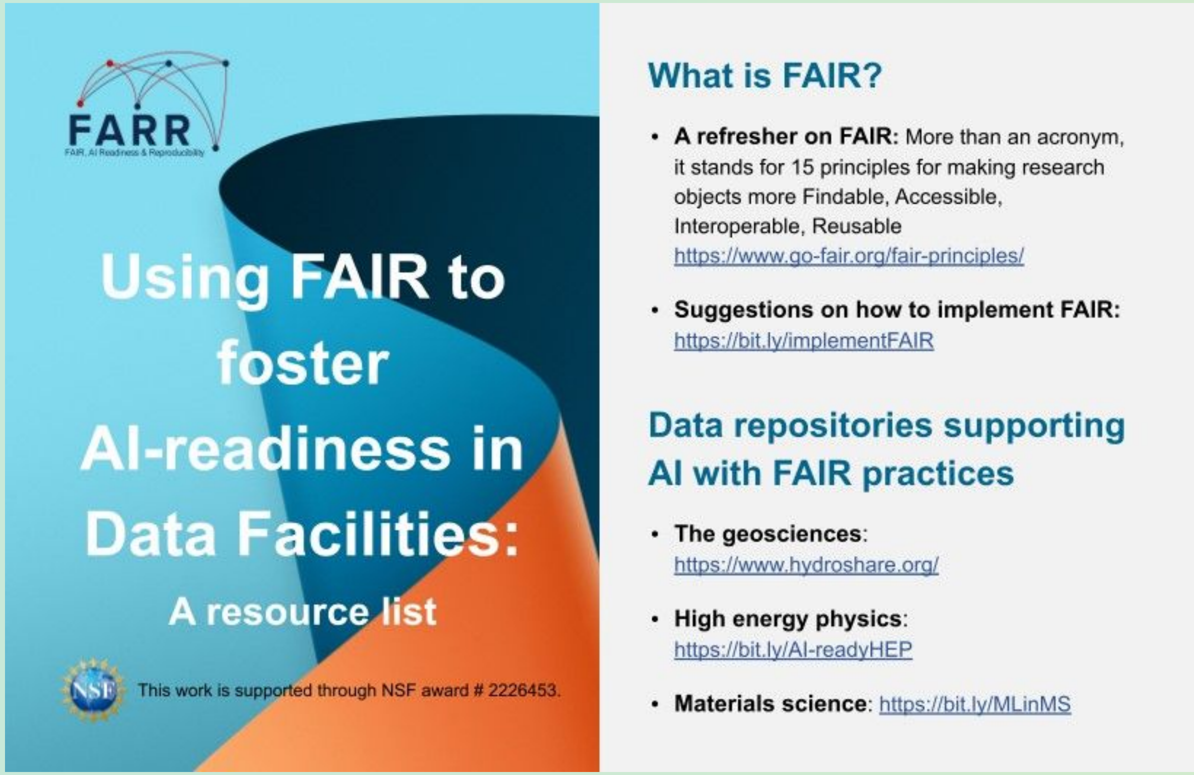
Kirkpatrick CR, Coakley K, Christopher J, Dutra I. Engaging with Researchers and Raising Awareness of FAIR and Open Science through the FAIR+ Implementation Survey Tool (FAIRIST). Data Science Journal. 2023; 22:32. Available from: https://datascience.codata.org/articles/10.5334/dsj-2023-032

FARR: FAIR in ML, AI Readiness, & Reproducibility Research Coordination Network

Ways to Get Involved

- **Input** on community needs, gaps & roadmap
- **Suggest use cases** and let us promote your project's use of AI and FARR-related practices
- Let us feature you in a **science story**

Contact:
https://www.farr-rcn.org/
community@farr-rcn.org



FARR
FAIR, AI Readiness & Reproducibility

Using FAIR to foster AI-readiness in Data Facilities:

A resource list

This work is supported through NSF award # 2226453.

What is FAIR?

- **A refresher on FAIR:** More than an acronym, it stands for 15 principles for making research objects more Findable, Accessible, Interoperable, Reusable https://www.go-fair.org/fair-principles/
- **Suggestions on how to implement FAIR:** https://bit.ly/implementFAIR

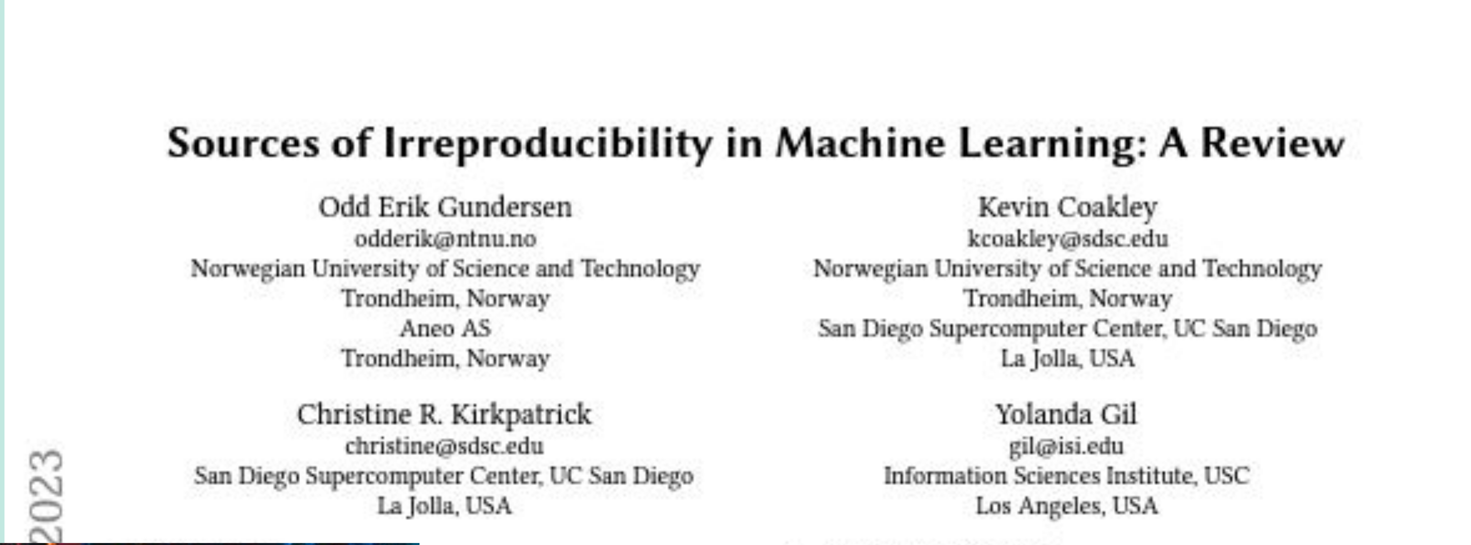
Data repositories supporting AI with FAIR practices

- **The geosciences**: https://www.hydroshare.org/
- **High energy physics**: https://bit.ly/AI-readyHEP
- **Materials science**: https://bit.ly/MLinMS

This work is supported through the NSF award #2226453.

Incorporating Knowledge from Papers

FARR
FAIR, AI Readiness & Reproducibility



Sources of Irreproducibility in Machine Learning: A Review

Odd Erik Gundersen
odderik@ntnu.no
Norwegian University of Science and Technology
Trondheim, Norway
Aneo AS
Trondheim, Norway

Kevin Coakley
kcoakley@sdsc.edu
Norwegian University of Science and Technology
Trondheim, Norway
San Diego Supercomputer Center, UC San Diego
La Jolla, USA

Christine R. Kirkpatrick
christine@sdsc.edu
San Diego Supercomputer Center, UC San Diego
La Jolla, USA

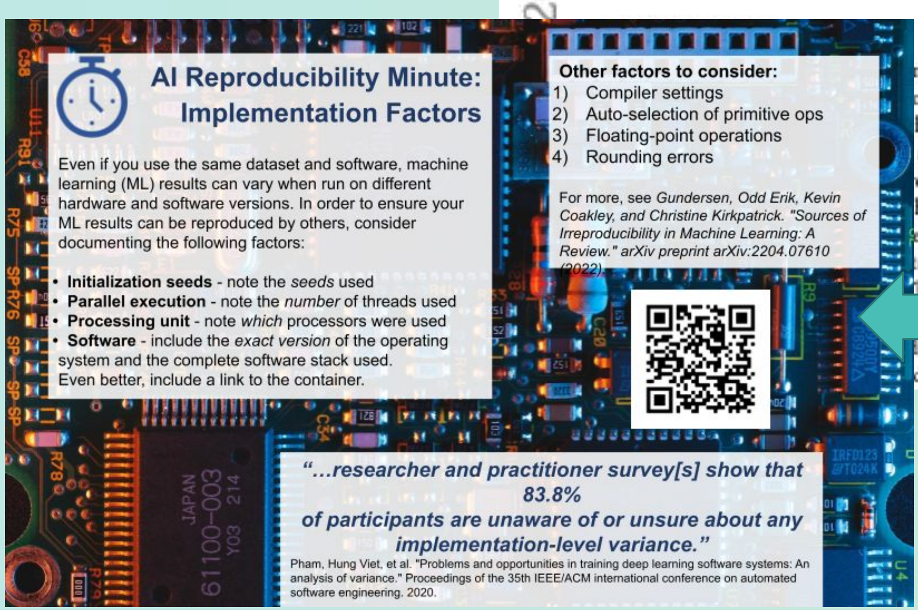
Yolanda Gil
gil@isi.edu
Information Sciences Institute, USC
Los Angeles, USA

2023

achine learning studies are irre-
ology and not properly account-
he algorithm themselves or their
as the main contributors to the
e exist no theoretical framework
hoices to potential effects on the
ework, it is much harder for practi-
e experiment results and describe
he lack of such a framework also
cally at-
ucibility experiments. **Objective:**
develop a framework that enable

1 INTRODUCTION

In recent years, many machine learning studies have shown to b challenging to reproduce. The areas of machine learning that have reported issues are very diverse and include forecasting [Makridakis et al. 2018], natural language processing [Belz et al. 2021a], generative adversarial networks [Lucic et al. 2018], deep reinforcement learning [Henderson et al. 2018], recommender systems [Dacrema et al. 2019], and image recognition [Bouthillier et al. 2019]. The authors above point to many methodological issues that are commonly found in machine learning research. Since applications of machine learning reach into many other fields [Gibney 2022], methodological shortcomings can have far reaching effects particularly in domains with high-stakes decisions such as medicine [Roberts et al. 2021; Varoquaux and Cheplygina 2022], social sciences [Kapoor and Narayanan 2022], psychology [Hullman et al. 2022] and many more [Raji et al. 2022].



AI Reproducibility Minute: Implementation Factors

Even if you use the same dataset and software, machine learning (ML) results can vary when run on different hardware and software versions. In order to ensure your ML results can be reproduced by others, consider documenting the following factors:

- **Initialization seeds** - note the *seeds* used
- **Parallel execution** - note the *number* of threads used
- **Processing unit** - note *which* processors were used
- **Software** - include the *exact version* of the operating system and the complete software stack used.

Even better, include a link to the container.

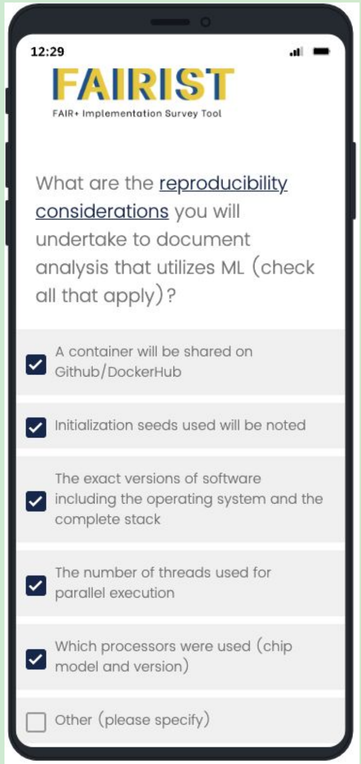
Other factors to consider:

1) Compiler settings
2) Auto-selection of primitive ops
3) Floating-point operations
4) Rounding errors

For more, see *Gundersen, Odd Erik, Kevin Coakley, and Christine Kirkpatrick. "Sources of Irreproducibility in Machine Learning: A Review." arXiv preprint arXiv:2204.07610 (2022).*

> ***"…researcher and practitioner survey[s] show that 83.8% of participants are unaware of or unsure about any implementation-level variance."***

Pham, Hung Viet, et al. "Problems and opportunities in training deep learning software systems: An analysis of variance." Proceedings of the 35th IEEE/ACM international conference on automated software engineering. 2020.



What are the reproducibility considerations you will undertake to document analysis that utilizes ML (check all that apply)?

☑ A container will be shared on Github/DockerHub

☑ Initialization seeds used will be noted

☑ The exact versions of software including the operating system and the complete stack

☑ The number of threads used for parallel execution

☑ Which processors were used (chip model and version)

□ Other (please specify)

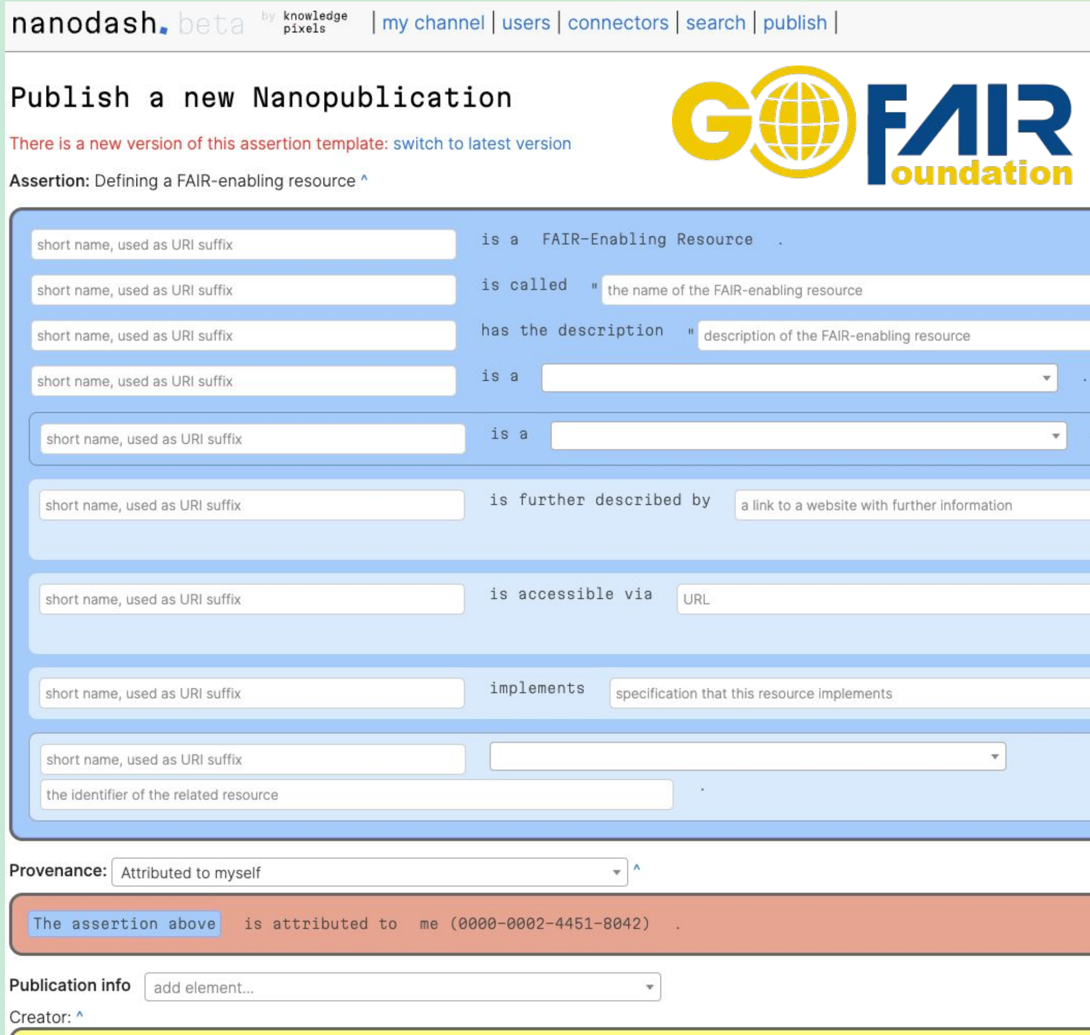
Intermediate step was to simplify into a postcard

Gundersen, O.E., Coakley, K., Kirkpatrick, C. and Gil, Y., 2022. Sources of irreproducibility in machine learning: A review. *arXiv preprint arXiv:2204.07610*.

Next Level: Nanopubs & FAIRIST

FAIR Connect
Empowering Data Stewardship

- Add additional FAIR, ethics, open science implementation options
  - Use knowledge from National Science Data Fabric catalog and FDO work
  - UNESCO open science recommendations & NASA TOPS
  - Emerging ethics work in CODATA, EU, domains
- Break down each practice into a chunk and publish as a nanopublication
- Granularity level of an RDF triple
  Subject+predicate+object
  Malaria is spread by mosquitoes
  Assign DOIs using DataCite-issued PIDs
- Nanopubs reviewed by peers (esp. data stewards)
- FAIRConnect for FAIR-enabling resource nanopubs → http://fairconnect.pro
- Could use threshold of endorsements for inclusion
- Tools could gather machine readable practices



nanodash. beta by knowledge pixels | my channel | users | connectors | search | publish |

Publish a new Nanopublication

GO FAIR Foundation

There is a new version of this assertion template: switch to latest version

Assertion: Defining a FAIR-enabling resource ^

short name, used as URI suffix is a FAIR-Enabling Resource .
short name, used as URI suffix is called " the name of the FAIR-enabling resource
short name, used as URI suffix has the description " description of the FAIR-enabling resource
short name, used as URI suffix is a
short name, used as URI suffix is a
short name, used as URI suffix is further described by a link to a website with further information
short name, used as URI suffix is accessible via URL
short name, used as URI suffix implements specification that this resource implements
short name, used as URI suffix
the identifier of the related resource .

Provenance: Attributed to myself ^

The assertion above is attributed to me (0000-0002-4451-8042) .

Publication info add element...

Creator: ^

Q&A

Thanks for listening!

Contact me at christine@sdsc.edu

https://www.linkedin.com/in/kirkpatrickchristine/

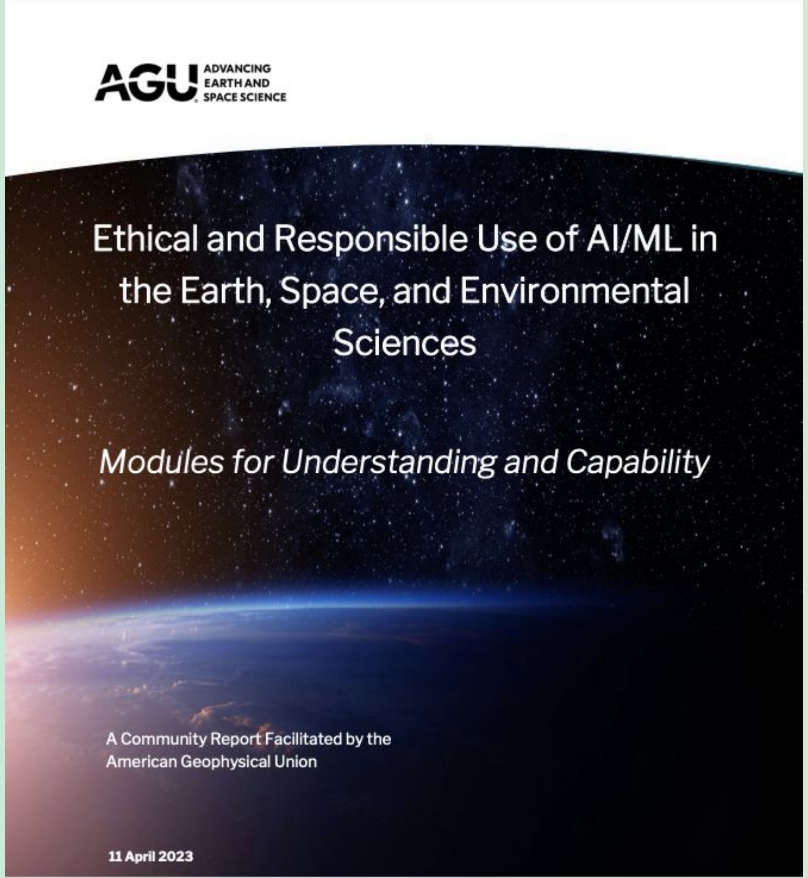
Ethics

- CARE Principles of Indigenous Data Governance CODATA Data Science Journal Editorial Policy:

Any use or consideration of Indigenous Knowledge should address The CARE Principles for Indigenous Data Governance and provide evidence of the care taken towards engagement with Indigenous communities including appropriate attribution, appropriate access, and **ideally Indigenous authorship**. Authors should include appropriate details of their perspective and background in the author description.

- AI Ethics
    - Ethical and Responsible Use of AI/ML (for Earth Sciences)
    - Emerging EU AI Regulation
        - Registering AI applications
        - Documenting adherence
    - US Executive Order on AI

Be
FAIR
Findable
Accessible
Interoperable
Reusable
and
CARE
Collective Benefit
Authority to Control
Responsibility
Ethics



AGU ADVANCING EARTH AND SPACE SCIENCE

Ethical and Responsible Use of AI/ML in the Earth, Space, and Environmental Sciences

Modules for Understanding and Capability

A Community Report Facilitated by the American Geophysical Union

11 April 2023

Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J. and Hudson, M., 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), p.43.DOI: https://doi.org/10.5334/dsj-2020-043

Shelley Stall, Guido Cervone, Caroline Coward, et al. Ethical and Responsible Use of AI/ML in the Earth, Space, and Environmental Sciences . *ESS Open Archive* . April 12, 2023.