Considerations for Data and AI Workflows

EIC Echelon 0-1 Mini-Workshop

8 October, 2025

Laura Biven

Chief Data Officer, Jefferson Lab

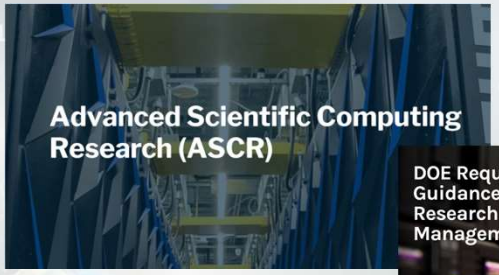
Jefferson Lab

U.S. Department of ENERGY

Laura Biven
Chief Data Officer

biven@jlab.org

Jefferson Lab



PuRe Data Resources

NCPI

AIM AHEAD

My interests:

- ***Enhancing innovation and integrity in science through the data lens***
- ***Building data infrastructure for creative, inquisitive research***

Considerations for Data Management and AI

1. Policies e.g. DOE O241.C
2. FAIR and AI-Ready Data
3. Reproducible, reusable, reinterpretable data analyses
4. Integrating with broader data / compute ecosystems e.g. HPDF, IRI, American Science Cloud
5. The Full Data Lifecycle view

DOE O241.C New Requirements

2023 DOE Public Access Plan

Publications

- Move from 12-month embargo to immediate access upon publication
- Continue to submit accepted manuscripts via E-Link, but earlier in reporting process
- Provide access through DOE's designated repository, DOE PAGES®
- Emphasize author deposits of accepted manuscripts (green OA) – DOE

Data

- Now Data Management and Sharing Plans (DMSPs)
- "Scientific Data" to validate and replicate research findings
- ★ Data underlying publications should be made available at time of publication
- ★ Timeline for sharing other scientific data
- ★ Repository selection should align with NSTC Desirable Characteristics of Data Repositories guidance

Persistent Identifiers

- Collect metadata associated with publications and data
- Metadata to include authors, affiliations and funding with associated PIDs, publication date, and PID for output
- ★ Instruct researchers to obtain a PID for themselves and use when publishing and reporting R&D outputs
- Researcher PIDs must meet common/core standards
- ? PIDs for awards

2023 DOE Public Access Plan: https://www.energy.gov/doe-public-access-plan

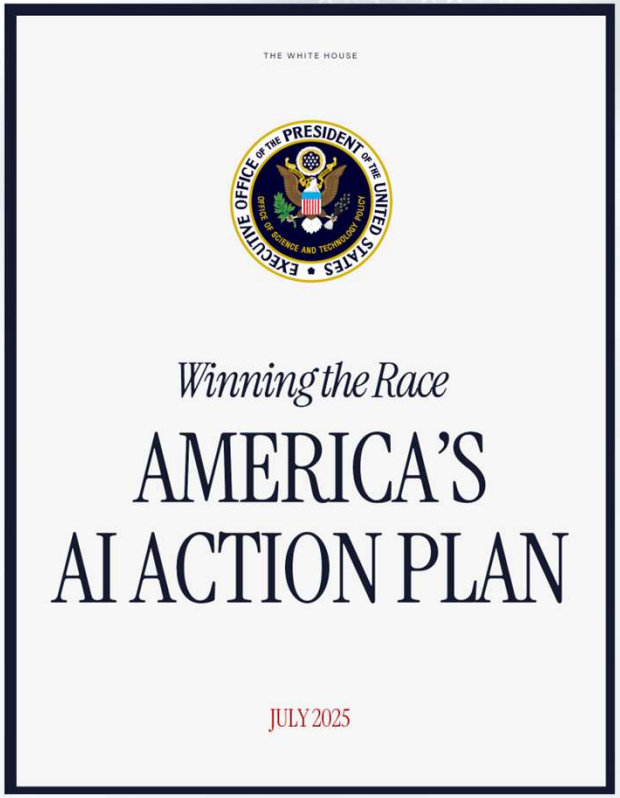
AI Action Plan

"Build World-Class Scientific Datasets

High-quality data has become a national strategic asset as governments pursue AI innovation goals and capitalize on the technology's economic benefits. Other countries, including our adversaries, have raced ahead of us in amassing vast troves of scientific data. The United States must lead the creation of the world's largest and highest quality AI-ready scientific datasets, while maintaining respect for individual rights and ensuring civil liberties, privacy, and confidentiality protections.

Recommended Policy Actions

Direct the National Science and Technology Council (NSTC) Machine Learning and AI Subcommittee to make recommendations on minimum data quality standards for the use of biological, materials science, chemical, physical, and other scientific data modalities in AI model training."



https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/

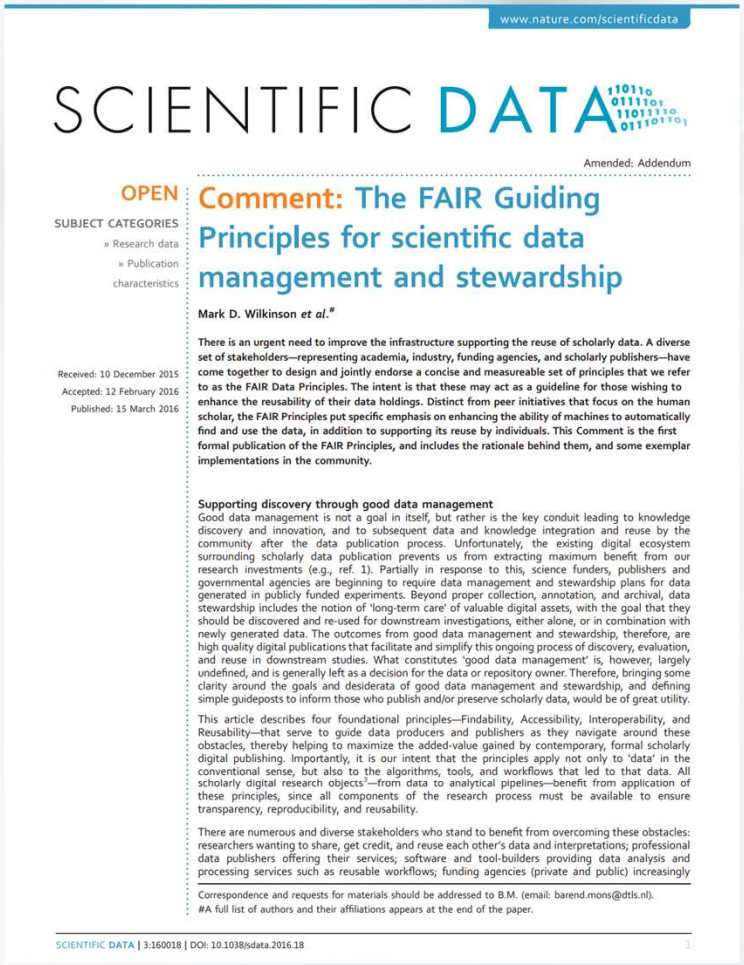
Jefferson Lab

DOE O241.X … Where is this going? (Laura's predictions)

- Increasing expectations for {data, code, documentation…} sharing
  - More data, more immediate sharing, more context,
  - Publications, data, and code are fully integrated, FAIR and AI-ready.
- Enhanced interplay between humans and AI
  - Assumptions and underlying theories (conceptual models) are described and shared along with research findings, data, and other research artifacts.
  - Increasing need for validation and resiliency for AI in the scientific process
- Tensions between protecting data and openness
- Emergence of global frameworks and standards for data and metadata

FAIR Principles

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18



www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

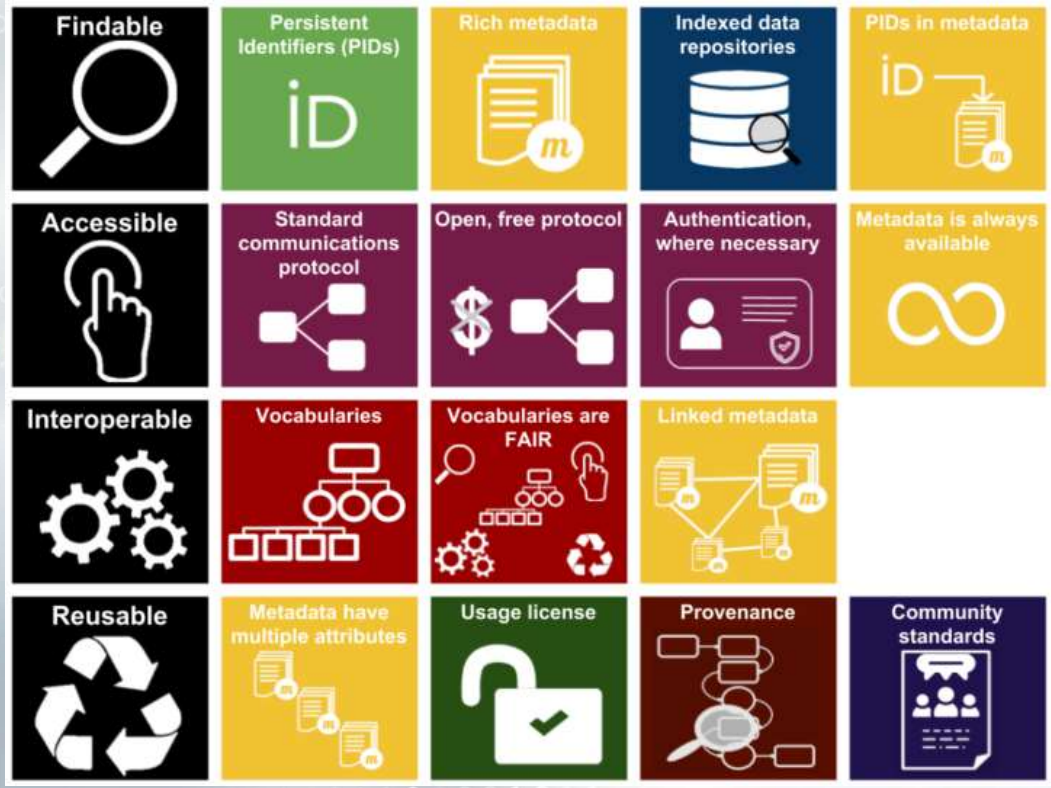
Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

FAIR Principles for Data



https://www.gofair.foundation/interpretation

FAIR data infographic (CC-BY except F.A.I.R logos CC-BY-SA by Sangya Pundir

FAIRification

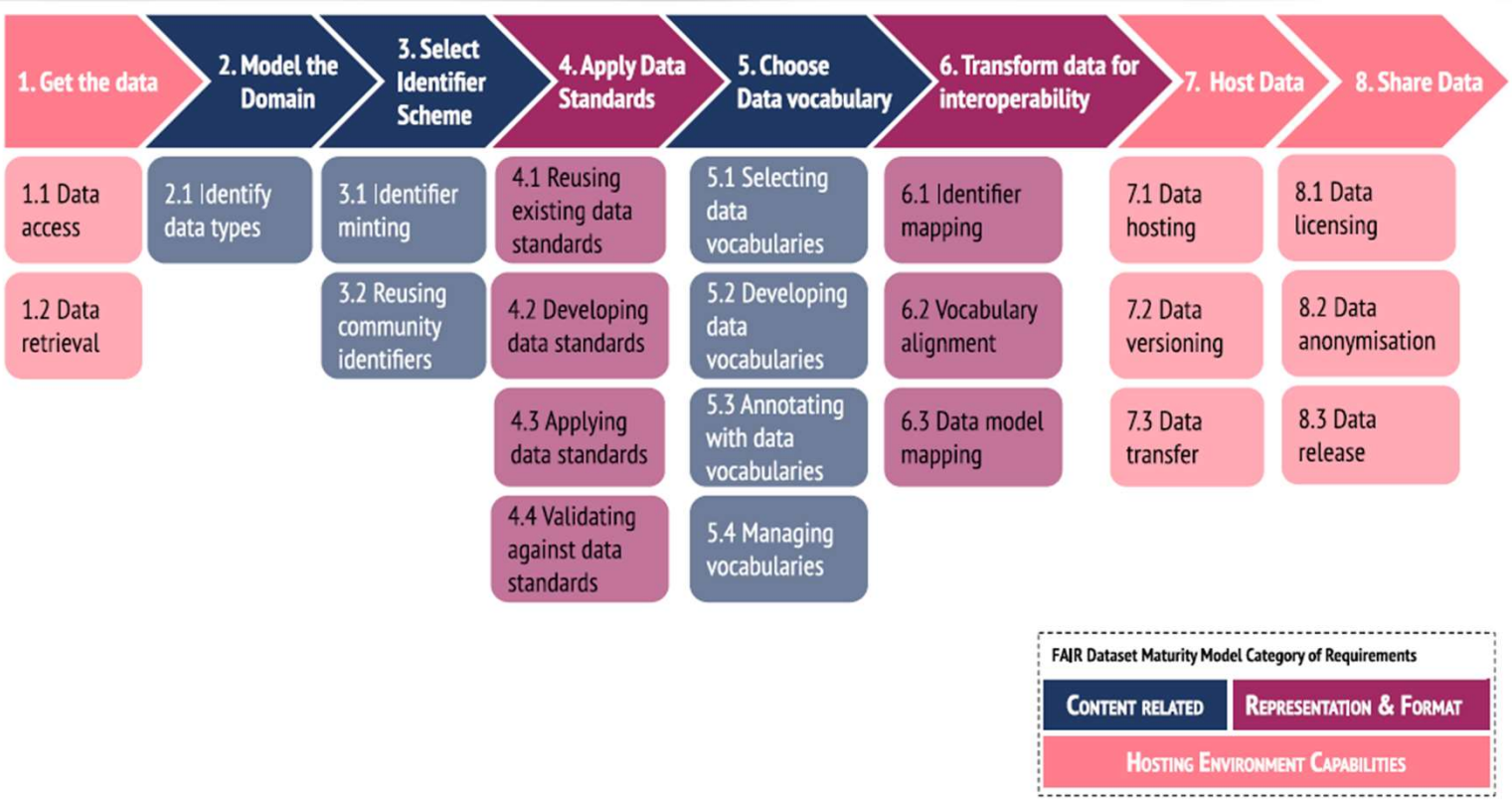


Fig. 4 The FAIRification template steps. Each step is colour-coded based on whether its implementation applies to data hosting, representation and format or data content. Each step is broken down into one or more sub-steps. More details can be found in Supplementary Table 2.

Welter, D., Juty, N., Rocca-Serra, P. *et al.* FAIR in action - a flexible framework to guide FAIRification. *Sci Data* **10**, 291 (2023). https://doi.org/10.1038/s41597-023-02167-2

FAIR4…

FAIR4Workflows: https://workflows.community/groups/fair/

FAIR4HEP: https://fair4hep.github.io/ , https://fairos-hep.org/

FAIR4RS: https://www.researchsoft.org/blog/2024-03/

FAIR4AI: https://doi.org/10.1038/s41597-023-02298-6

FAIR4HPC: https://hpc-fair.github.io/ , https://doi.org/10.1145/3708035.3736097

FAIR training Materials https://doi.org/10.1371/journal.pcbi.1007854

FAIRDO – Digital Objects https://fairdo.org/

FAIR Principles for Research Hardware https://www.rd-alliance.org/groups/fair-principles-research-hardware

Desirable Characteristics for Repositories https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf

Public Law No: 119-21

SEC. 50404. TRANSFORMATIONAL ARTIFICIAL INTELLIGENCE MODELS.

AMERICAN SCIENCE CLOUD.—The term ‘‘American science cloud’’ means a system of United States government, academic, and private sector programs and infrastructures utilizing cloud computing technologies to facilitate and support scientific research, data sharing, and computational analysis across various disciplines while ensuring compliance with applicable legal, regulatory, and privacy standards.

TRANSFORMATIONAL MODELS.—The Secretary of Energy shall—

(1) mobilize National Laboratories to partner with industry sectors within the United States to curate the scientific data of the Department of Energy across the National Laboratory complex so that the data is structured, cleaned, and preprocessed in a way that makes it suitable for use in artificial intelligence and machine learning models; and

(2) initiate seed efforts for self-improving artificial intelligence models for science and engineering powered by the data described in paragraph (1).

https://www.congress.gov/bill/119th-congress/house-bill/1

The American Science Cloud at DOE

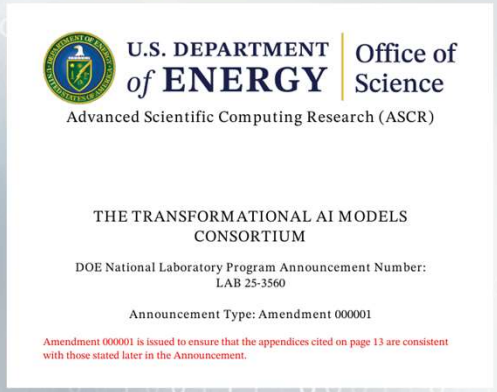

U.S. DEPARTMENT of ENERGY | Office of Science

Advanced Scientific Computing Research (ASCR)

THE AMERICAN SCIENCE CLOUD (AmSC)

DOE National Laboratory Program Announcement Number: LAB 25-3555

Announcement Type: Initial


U.S. DEPARTMENT of ENERGY | Office of Science

Advanced Scientific Computing Research (ASCR)

THE TRANSFORMATIONAL AI MODELS CONSORTIUM

DOE National Laboratory Program Announcement Number: LAB 25-3560

Announcement Type: Amendment 000001

Amendment 000001 is issued to ensure that the appendices cited on page 13 are consistent with those stated later in the Announcement.

"AI is the next Manhattan Project. AI technology will define the future of the world, and it is essential that the U.S. leads in the development of this technology. DOE has a significant role to play in driving AI innovation for scientific discovery, energy innovation, and national security."

- Secretary of Energy, Chris Wright (May 2025)
https://www.appropriations.senate.gov/imo/media/doc/secretary_wright_testimony.pdf

National-scale data and computing infrastructures

U.S. DEPARTMENT of ENERGY | Office of Science

INTEGRATED RESEARCH INFRASTRUCTURE

PROGRAM

SEC. 50404. TRANSFORMATIONAL ARTIFICIAL INTELLIGENCE MODELS. 15 USC 9461 note.

(a) DEFINITIONS.—In this section:

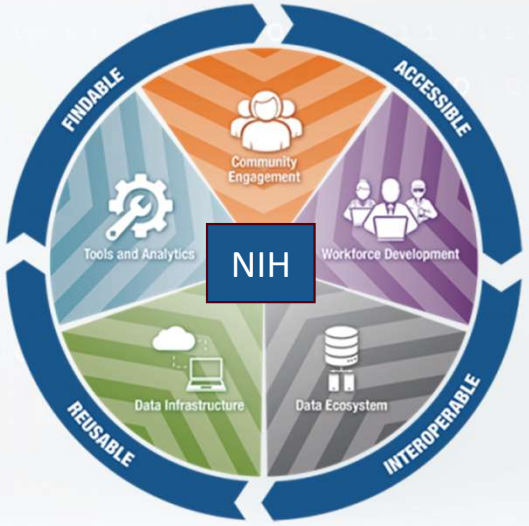
(1) AMERICAN SCIENCE CLOUD.—The term "American science cloud" means a system of United States government, academic, and private sector programs and infrastructures utilizing cloud computing technologies to facilitate and support scientific research, data sharing, and computational analysis across various disciplines while ensuring compliance with applicable legal, regulatory, and privacy standards.

(2) ARTIFICIAL INTELLIGENCE.—The term "artificial intelligence" has the meaning given the term in section 5002 of the National Artificial Intelligence Initiative Act of 2020 (15 U.S.C. 9401).

High Performance Data Facility: Supporting the Data Life Cycle

OUR MISSION: To enable and accelerate scientific discovery by delivering state-of-the-art data management infrastructure, capabilities, and tools to the nation's research communities.

NAIRR Pilot National Artificial Intelligence Research Resource Pilot



Considering the Full Data Lifecycle

- Identify use cases and users broadly
  - For your own collaborations: reproducibility, reusability, reinterpretability, automated generation of analyses.
  - For other nuclear scientists, future nuclear scientists, students, AI agents, AI researchers, public,…
  - For cross experiment analysis
- Define governance and sharing rules in alignment with policies and expectations.
- Steward the full scientific record, considering sustainability: publication, code, data, provenance, documentation (datasheets, model cards)
- Leverage standards and infrastructure from national-scale infrastructures
- Continue detailed planning for workflows through E3 and beyond.
- Everything is data for AI – test frequently
  - Opportunity to consider the full corpus of data, information, and knowledge for AI

➢ Importance of metadata, provenance, documentation (and which is which), curated repos,

➢ Importance of persistent identifiers for data, code, people (ORCID), institutions, detector components

➢ Importance of reusable, modular, interoperable software and workflow components.

➢ Importance of standards and interoperability

Jefferson Lab

Thank you

U.S. Department of ENERGY