Data Lifecycle and FAIR Principles for ePIC: Priorities and Next Steps

AI in ePIC

23 January, 2026

Laura Biven

Chief Data Officer, JLAB

Jefferson Lab

U.S. Department of ENERGY

1. The Chief Data Office
2. External factors (forcing functions and boundary conditions)
3. FAIR and AI Ready Data & Reproducible, Reusable, Reinterpretable Data Analyses
4. Chief Data Office efforts at JLab

JLAB Chief Data Office

https://www.jlab.org/about/leadership/cdo_office

Laura Biven
Chief Data Officer

My interests:

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

Diana McSpadden
Data Scientist – Data Steward

My interests:

- ***Data stewardship for scientific data***
- ***ML and uncertainty quantification for coastal flood management***

The Genesis Mission: Prioritizing American Science and Technology Leadership

"The Genesis Mission will mobilize the Department of Energy's 17 National Laboratories, industry, and academia to build an integrated discovery platform—The Genesis Science and Security Platform. The platform will connect the world's fastest supercomputers, AI systems, and next-generation quantum computers with the most exquisite scientific instruments and data in the Nation…

This mission embodies our ambition to **dramatically accelerate scientific discovery** and to significantly increase the productivity and impact of R&D in the United States, which we aim to **double within a decade**."

- Under Secretary for Science, Darío Gil (December 10, 2025)

From testimony before the House Committee on Science, Space, and Technology

The AI Mission Activities from DOE-SC

Model Teams	Public/private partnerships to **develop self-improving AI models** across various science and engineering domains as part of the consortium
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	We propose establishing the Transformational AI Models Consortium (ModCon) to **lead the data and model-building activities** outlined in Section 50404… (1) establishing a consortium to **facilitate the creation of and to host MTs**…, (2) standing up a set of crosscutting, domain-independent services to provide the **base engine for building advanced AI models**; and (3) **convening partners** from industry, academia, and internationally to accelerate the development and adoption of AI
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	The AmSC will serve as the **enabling software and hardware infrastructure** for DOE's AI data and model development efforts in furtherance of SC's mission and in fulfillment of Section 50404 of the OBBB Act.

HPDF
High Performance Data Facility

July 2025: 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/

2023: 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***

Layers of data management

- Persistent Identifiers
- Metadata
- Data Documentation
- Governance
- Reusability

Thinking about interoperable data

- ORCID(s) for individual researchers
- ROR ID(s) for research organizations
- Grant ID(s) for funding sources
- DOI(s) for Dataset(s)
- DOI(s) for Datasheet(s)
- DOIs / citations for software and code
- DOI(s) for manuscripts

- Allows for distinct authors, research orgs, resources, funding sources, etc for each atom of the scientific record.
- **Allows for linking the elements of the scientific record.**

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

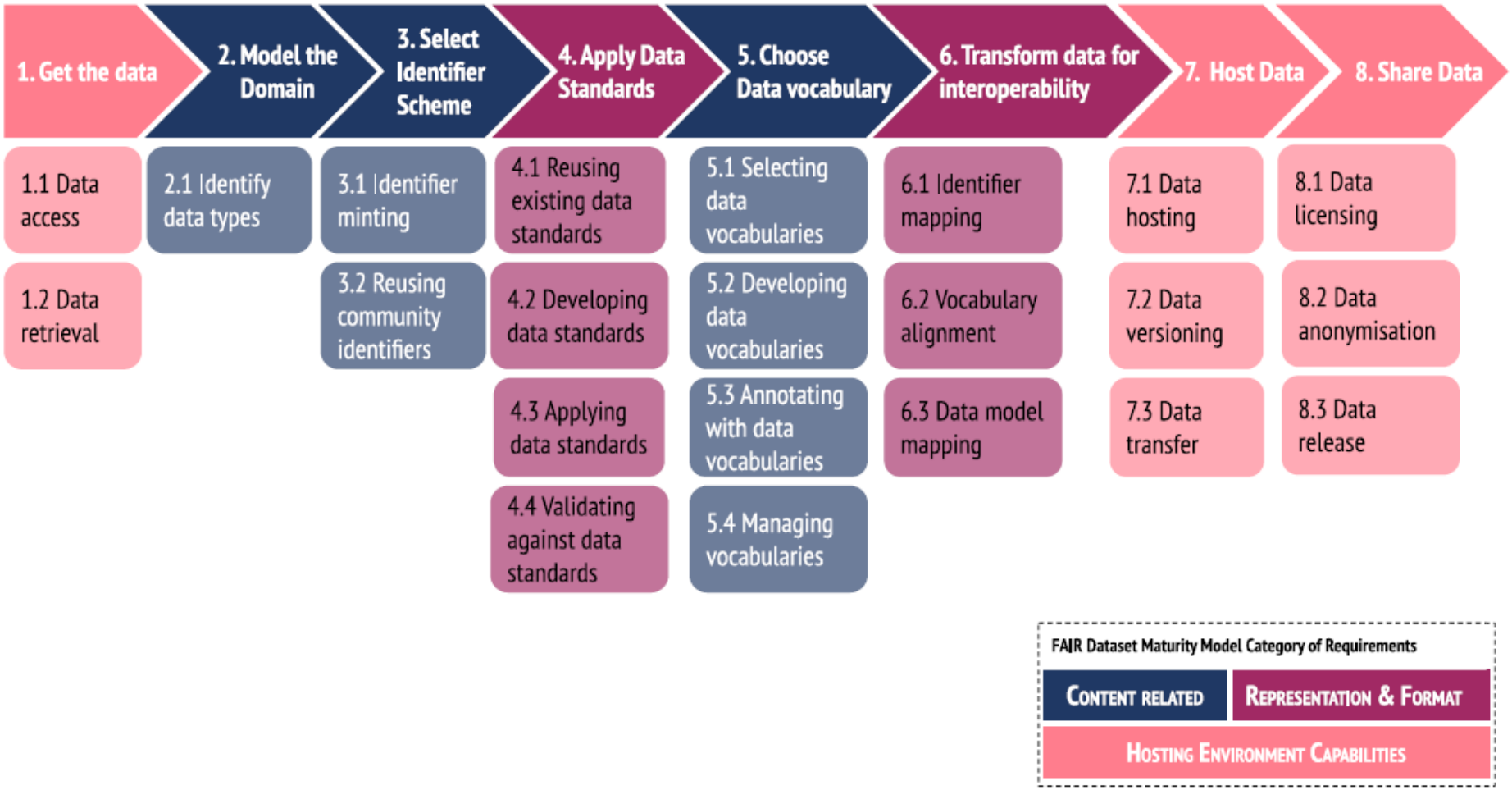


Fig. 4 The FAIRification template steps. Each step is colour-coded based on whether its implementation applies to data hosting, representation and format or data content. Each step is broken down into one or more sub-steps. More details can be found in Supplementary Table 2.

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

Views of AI-Ready Data

Many labs have developed AI-readiness eval tools. Example:
Hiniduma, K., Byna, S., Bez, J. L., & Madduri, R. (2024, July). Ai data readiness inspector (aidrin) for quantitative assessment of data readiness for ai. In *Proceedings of the 36th International Conference on Scientific and Statistical Database Management* (pp. 1-12). https://doi.org/10.1145/3676288.3676296

Model-centric

- Can I train a model with this data?
- Clean data
- Consistent units
- No manual preprocessing required

Task-contextual

- Ready for what type of AI?
- E.g., Classification, surrogate modeling, foundation models
- There is no universal AI-ready standard, only AI-ready for X.

Provenance-aware

- Can I reproduce this workflow?
- Traceability, auditability, reproducibility
- Lineage, transformations, and versions are documented
- Path from raw to the intermediate steps is clear.

Knowledge-enriched

- Data + meaning + structure
- Data coupled with context/explicit knowledge
- Ontologies, knowledge graphs, and semantic relationships
- Can enable reasoning, not just prediction

FAIR + Machine actionable

- Data is FAIR by design
- Persistent IDs link people, projects, datasets, code
- Shared vocabularies & ontologies align meaning across terms
- Rich structured metadata is complete and consistent
- Enables cross-lab AI use
- Not always ready to train without other views in place (e.g., **model-centric** and **provenance**)

Hiniduma, Kaveen, et al. "Data Readiness for AI: A 360-Degree Survey," ACM Comput. Surv., vol. 57, no. 9, Apr. 2025. DOI:10.1145/3722214

Brewer, W., Widener, P., Anantharaj, V., Wang, F., Beck, T., Shankar, A., & Oral, S. (2025, September). Data Readiness for Scientific AI at Scale. In Workshop Proceedings of the 54th International Conference on Parallel Processing (pp. 18-24). DOI:10.1145/3750720.3757282

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.

Dario Gil's vision for "the world's largest and highest-quality scientific data sets to train the next generation of AI systems" (https://www.energy.gov/science/articles/under-secretary-gils-letter-community).

- Tensions between protecting data and openness
- 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
- Emergence of global frameworks and standards for data and metadata
  - Shameless plug for FDO Conference: https://fairdo.org/fdo-conference-2026-registration/

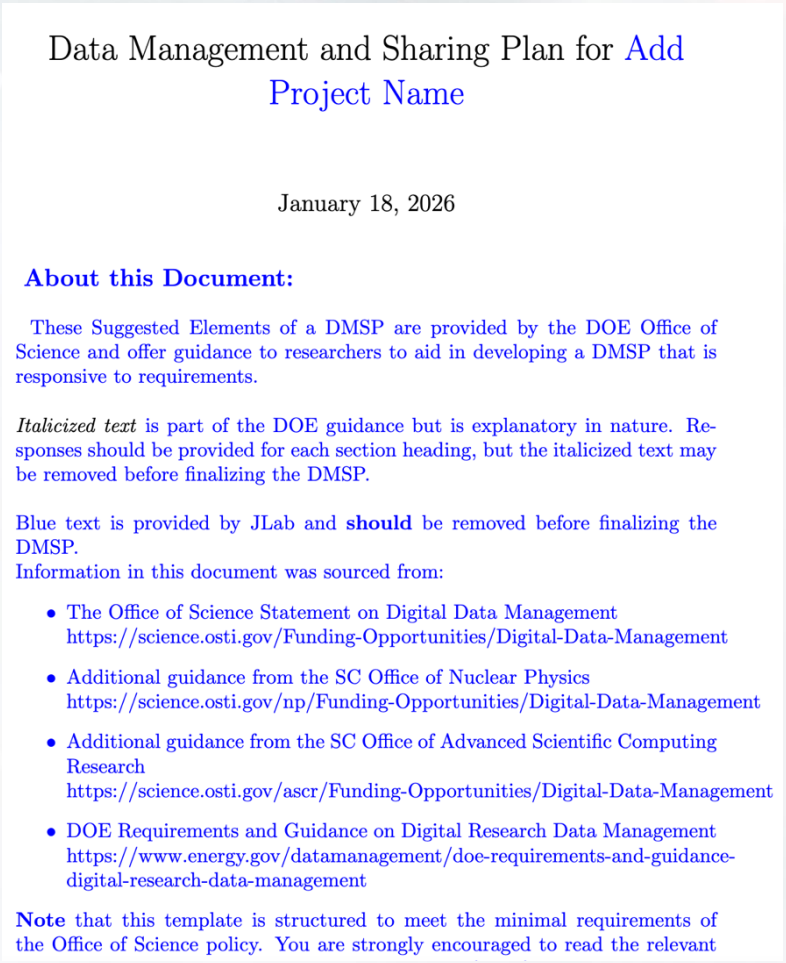
Chief Data Office efforts at JLab

The AI Mission Activities from DOE-SC – JLab Awards $7.15M total

Model Teams	Seed Model Teams	
		MT: MOAT: Multi-Office Particle Accelerator Team (LBNL) - Tennant
		MT: Q2C: From Quarks to Cosmos through the Lens of AI (ANL) - Sato
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	**ModCon (ANL)** Biven, Britton, McSpadden, Sawatzky, Goldenberg, Schuchman	**NPDaPP: NARAD: NP AI Ready Accelerator Data (JLab)** - Satogata
		NPDaPP: Preparing QCD Data for Foundation Models (BNL) - Biven
		Biven Co-leads the Data Broker and Standards Team
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	**AmSC (ORNL)** Baldin, Hayes, Hess, Gyurjyan, Mei, Singh	**IP: HPDF** - Heyes
		IP: BES/HEP/NP SUF (LBNL) - Lawrence
		IP: Foundational Infrastructure for QCD Science - Edwards
		IP: HAIDIS (Hardware-enable AI Distributed Inverse Solver)- Baldin

JLab Data Management and Sharing Plan Template

- Chief Data Office and Record Management have developed template and guidance for completing Data Management and Sharing Plans
  - Structured to meet the minimal requirements of the Office of Science policy:
    - Data types, sources, and standards
    - Related tools, software and code
    - Data access and reuse considerations
    - Data security, preservation, and sharing
    - Oversight of data management and sharing
    - Estimating the cost of implementing a DMSP
- Chief Data Office requesting feedback



Data Management and Sharing Plan for Add Project Name

January 18, 2026

About this Document:

These Suggested Elements of a DMSP are provided by the DOE Office of Science and offer guidance to researchers to aid in developing a DMSP that is responsive to requirements.

Italicized text is part of the DOE guidance but is explanatory in nature. Responses should be provided for each section heading, but the italicized text may be removed before finalizing the DMSP.

Blue text is provided by JLab and **should** be removed before finalizing the DMSP.
Information in this document was sourced from:

- The Office of Science Statement on Digital Data Management https://science.osti.gov/Funding-Opportunities/Digital-Data-Management
- Additional guidance from the SC Office of Nuclear Physics https://science.osti.gov/np/Funding-Opportunities/Digital-Data-Management
- Additional guidance from the SC Office of Advanced Scientific Computing Research https://science.osti.gov/ascr/Funding-Opportunities/Digital-Data-Management
- DOE Requirements and Guidance on Digital Research Data Management https://www.energy.gov/datamanagement/doe-requirements-and-guidance-digital-research-data-management

Note that this template is structured to meet the minimal requirements of the Office of Science policy. You are strongly encouraged to read the relevant

Context for Data

Jefferson Lab

metadata

- Highly structured (JSON, XML, RDF)
- Often community-specific
- Enables data to be found, indexed, managed, understood by software systems/AI; supports reuse
- Adjacent to files in repo, or published separately
- Unpublished & published datasets

Datacard / README

- Includes an overview and context, organization & file structure, data details, methodological info, access & reuse
- Human & AI-readable
- Adjacent to files in data repo
- Published & unpublished datasets

Datasheet

- Comprehensive documentation of creation, scope, intended use, limitations, & ethical/technical considerations
- Capture tacit information & tribal knowledge
- Enables frontier models to leverage rich, unstructured dataset information
- For published datasets

Context for Data

Jefferson Lab

metadata

- Highly structured (JSON, XML, RDF)
- Often co
- Enables indexed, understo systems/ reuse
- Adjacent or publis
- Unpublished & published datasets

Datacard / README

- Includes an overview and context, organization &

Datasheet

- Comprehensive documentation of , intended , & al
- formation dge
- r models to
- ataset information
- For published datasets

Genesis Mission is developing minimum metadata requirements and templates for Datacard and Datasheets

Genesis Mission is developing demonstrations and schema for information architecture

Governance

We are working with a few collaborations and divisions at JLab on data sharing rules and processes.

Very preliminary discussions at the moment.

We welcome engagement with other groups

Architecting the information landscape

March 25-27, 2026
Vienna, Austria

https://fairdo.org/fdo-conference-2026-registration/

Final thoughts on how we can help you

- The Chief Data Office (Laura and Diana) is here to support you in sharing your data effectively.

- The Chief Data Office is available to assist with guidance and assistance through the data sharing process.

- We WANT to collaborate with you!

Jefferson Lab

Thank you

U.S. Department of ENERGY