FAIR Software Meets FAIR Data

EPIC User Learning Workshop: Discoverability and Reusability for the preTDR and Beyond

2026-04-29

Diana McSpadden

Data Scientist–Data Steward, JLAB

Jefferson Lab

U.S. Department of ENERGY

1. Motivating factors
2. FAIR data and FAIR research software
3. Data lifecycle management and datacards
4. The Chief Data Office at JLab

Motivating Factors

Requirements, Initiatives, Your Science!

https://xkcd.com/3232/#:~:text=https%3A//xkcd.com/3232/

Some Terminology & DOE Public Access Plan

Data Sharing

Data sharing means **making data available to people other than those who have generated them**. Examples of data sharing range from bilateral communications with colleagues, to providing free, unrestricted access to the public through, for example, a web-based platform.

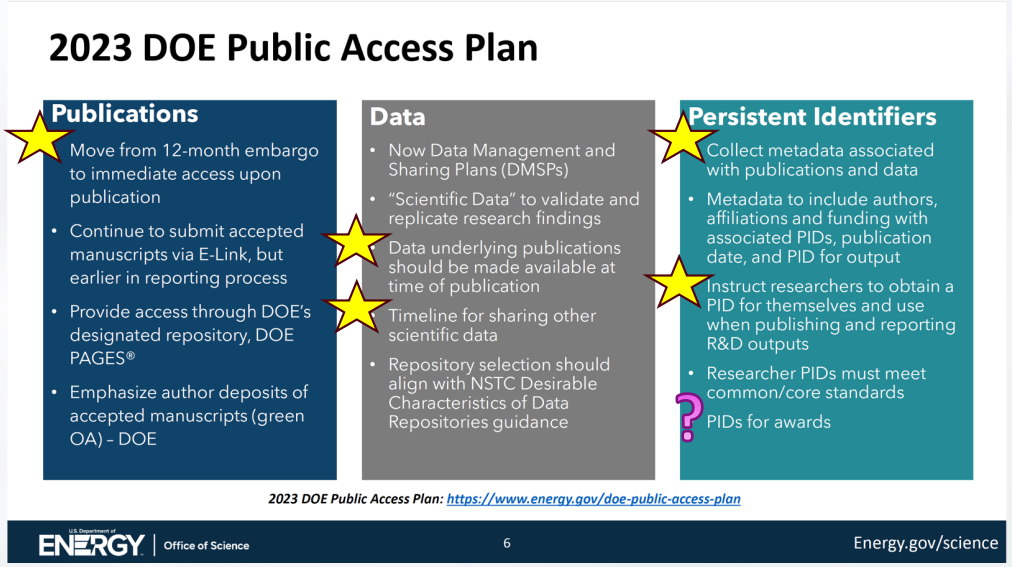
Appropriate Sharing

The term “appropriate” is used to signal that public access to Federally-funded research results and data should be maximized in a manner that **protects** confidentiality, privacy, business confidential information, and security, avoids negative impact on intellectual property rights, innovation, program and operational improvements, and U.S. competitiveness, and **preserves the balance between the relative value of long-term preservation and access and the associated cost and administrative burden**.

(https://www.energy.gov/datamanagement/glossary)

Open Data

Open data is data that can **be freely used, re-used and redistributed by anyone** - subject only, at most, to the requirement to attribute and sharealike. (https://opendatahandbook.org/guide/en/what-is-open-data/)

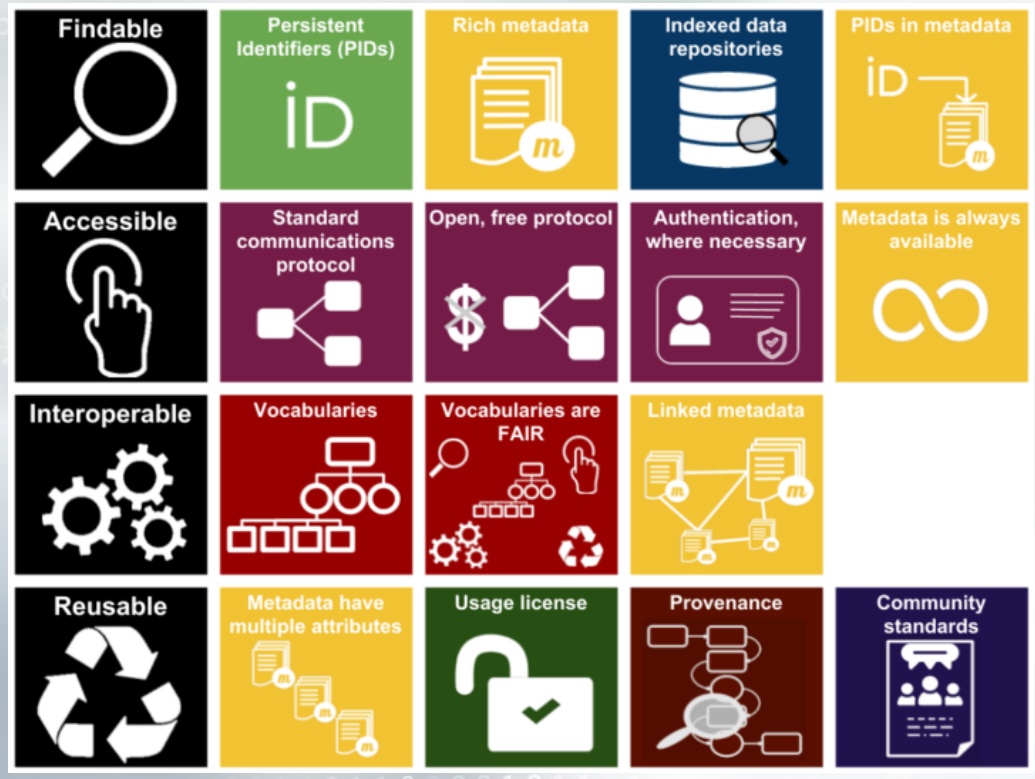


Gaps

- Each community needs to establish norms for "**data underlying a publication**" and revisit these as technologies and expectations change.
- Each community needs to establish norms for **data sharing timelines** and revisit these as technologies and expectations change.
- **Opportunity:** Sharing scientific workflows together with the assets that enable reuse (e.g., datasets, computational workflows, software, and models)?

FAIR Data Meets FAIR Research Software

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)

Why FAIR Research Software?

Jefferson Lab

Lamprecht, A. L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., ... & Capella-Gutierrez, S. (2020). Towards FAIR principles for research software. *Data Science*, *3*(1), 37-59. https://doi.org/10.3233/DS-190026.

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

FAIR for data is not sufficient → FAIR for software

Modern science is software-defined:

- Most results depend on complex, evolving codebases [1]
- Software encodes **methods, assumptions, and workflows** – enabling reproducibility

FAIR4RS & FAIR

Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., Honeyman, T., Struck, A., Lee, A., Loewe, A., van Werkhoven, B., Jones, C., Garijo, D., Plomp, E., Genova, F., … RDA FAIR4RS WG. (2022). FAIR Principles for Research Software (FAIR4RS Principles) (1.0). Zenodo. https://doi.org/10.15497/RDA00068

Barker, M., Chue Hong, N. P., Katz, D. S., Lamprecht, A.-L., Martinez, C., Psomopoulos, F., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., & Honeyman, T. (2022).
FAIR principles for research software (FAIR4RS principles). Scientific Data, 9, 622. https://doi.org/10.1038/s41597-022-01710-x

Software introduces executability, dependencies, and relationships.

FAIR4RS Principles
Findable: Software, and its associated metadata, is easy for both humans and machines to find. E.g., **F4: Metadata are FAIR, searchable and indexable.**
Accessible: Software, and its metadata, is retrievable via standardized protocols. E.g., **A1: Software is retrievable by its identifier using a standardized communications protocol.**
Interoperable: Software interoperates with other software by exchanging data and/or metadata, and through interaction via APIs, described through standards. E.g., **I1: Software reads, writes, and exchanges data in a way that meets domain-relevant community standards.**
Reusable: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software). E.g., **R2: Software includes *qualified references* to other software.**

OSTI, DataCite & Proposed Relationships: depends on, extends, implements, wraps, is compatible with, was derived from, consumes, generates, was trained on, validates against, requires metadata from, is compatible with

FAIR for Research Software

Git Repos are not enough

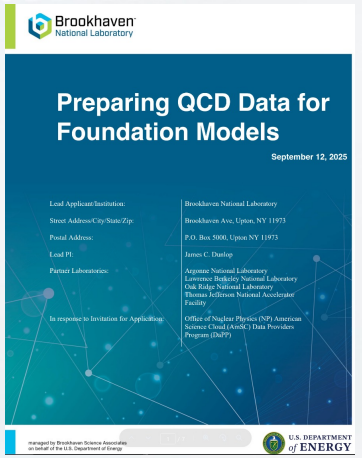
1. Missing Metadata
   - Git = code storage
   - FAIR = metadata-driven
2. Missing Relationships
   - Git does not encode:
     - “this model depends on that dataset”
     - “this workflow produces that output”
3. No Preservation Guarantee
   - GitHub ≠ archive
   - FAIR requires long-term stewardship
4. Execution ≠ Reproducibility or Reusability
   - “Code runs on my machine” problem

From code repos to FAIR ecosystem

1. Git (baseline)
   - Code + version control
2. Metadata
   - CITATION.cff, codemeta.json, DOI (Zendodo, Software Heritage)
3. **Reusability**
   - Containers, environments, snakemake, MLFlow experiment tracking
4. Interoperability
   - Schemas, APIs, shared data models, e.g., LinkML, JSON Schemas
5. Standardized APIs
6. **Relationships**
   - IDs
   - Graph/ontology layer (e.g., RDF/PROV-O)

Implementation for the EIC community

- Work with your lab's OSTI POC
  - ensures that the data and software products we produce are registered, discoverable, and properly documented.
  - OSTI is not a repository
    - can mint DOIs
    - Can support metadata curation
    - Supports findability
- Work with your tech transfer office
  - IP considerations
  - Clear licenses and reuse



Data lifecycle management & datacards

Jefferson Lab

Data lifecycle management for sharing and reuse

- Data Management Planning
- Persistent Identifiers
- **Metadata & Data Documentation – datacards!**
- Governance
- Data Sharing
- **Provenance and Analysis Preservation**
- **Reuse**

Metadata and Documentation

metadata

- Highly structured (JSON, YAML, RDF)
- Often community-specific
- Enables data to be found, indexed, managed and understood by software systems/AI; supports reuse
- Adjacent to files in a repository, or can be published separately
- For unpublished and published datasets

datacard / README

- May include structured metadata
- Includes an overview and the context, organization and files, structure, data details, methodological information; can capture community information; support access, interoperability and reuse
- Human and AI-readable
- May be adjacent to files in the repo, a separate webpage, or published separately
- For unpublished and published datasets

datasheet

- Comprehensive documentation of creation, scope, intended use, limitations, ethical/technical considerations
- Captures tacit information and community knowledge
- Enables leveraging of rich, unstructured dataset information
- For elite, published datasets

What is a datacard?

- A structured metadata document describing a dataset for both humans and machines: yaml + md
- A datacard answers: What is the data? Where did it come from? Who created it? How can it be accessed and used?

Three readiness levels define what to fill in:

Level	Profile	Goal	Who needs it
1	Discoverable	Enough metadata to find and identify the dataset	All datasets
2	Interoperable and Reusable	Accessible, governed, licensed and documented for sharing datasets outside the data creators	Shared datasets
3	AI-/Workflow-Ready & Trustworthy	Provenance-aware, integrity-supported, semantically clear	AI/ML, agentic, & reusable workflows.

Datacard template

The template annotates every field:

```yaml
# --- Identification ----------------------------------------------
identification:
  name: "${DATASET_NAME}"                    # [core] Single human-readable name for this dataset.
                                             # Use the same name in the datacard filename.
                                             # If this datacard covers a collection, provide the collection name.
  project: "${PROJECT_NAME}"                 # [core] Genesis project or sub-project this dataset belongs to.
                                             # e.g., genesis | genesis-fusion | genesis-lightsource
  version: "1.0"                             # [core] Dataset version using semantic versioning: MAJOR.MINOR.PATCH
```

Data quality is explicit:

```yaml
# --- Data Quality ------------------------------------------------
# [ai_ready] Be specific — vague entries reduce trust and reuse.
data_quality:
  completeness: __DESCRIPTION__              # [ai_ready] e.g., "All detector channels present; 2% of timesteps
                                             # missing due to instrument downtime on 2023-04-12"
  known_issues: __DESCRIPTION__              # [ai_ready] e.g., "Sensor drift observed after 2023-06-01T12:00:00Z"
  validation_methods: __DESC__               # [ai_ready] e.g., "Cross-validated against NIST SRM 640f"
  noise_characteristics: __DESC__            # [if_applicable]
  uncertainty_notes: __NOTES__               # [if_applicable] e.g., "Measurement uncertainty ±0.5% (k=2) per ISO/IEC Guide 98-3"
  missing_data_codes: []                     # [if_applicable]
```

Datacard template

Workflow state and release status travel together:

```yaml
# --- Release Status ------------------------------------------
release_status: ${STATUS}                    # [core] Current publication and governance state of this dataset.
                                             # See NOTE ON WORKFLOW STATE vs. RELEASE STATUS in the header
                                             # for expected alignment with workflow.state.
                                             #   draft         - work in progress; not ready for sharing
                                             #   under_review  - submitted for formal review
                                             #   approved      - review complete; cleared for release
                                             #   published     - publicly released and accessible
                                             #   deprecated    - superseded or retired; no longer recommended for use

# --- Workflow & Lifecycle ------------------------------------
# [core] Describes the technical and processing lifecycle position
# of the dataset. See NOTE ON WORKFLOW STATE vs. RELEASE STATUS
# in the header for expected alignment with release_status.
workflow:
  state: ${STATE}                            # [core] Current lifecycle position:
                                             #   raw        - data as collected; no processing applied
                                             #   processing - actively being cleaned, transformed, or reduced
                                             #   qa         - undergoing quality assurance or validation
                                             #   analysis   - in active scientific analysis
                                             #   review     - under formal review (security, export, IRB, etc.)
                                             #   embargo    - complete but intentionally withheld from release
                                             #   published  - publicly released
                                             #   archived   - preserved; no longer actively maintained
  is_intermediate: __BOOL__                  # [if_applicable] true if this is an intermediate processing artifact
                                             # rather than a final deliverable. false if this is final.
  pipeline_stage: __STAGE__                  # [if_applicable] Freetext position in processing pipeline.
                                             # e.g., "post-detector, pre-reconstruction"
                                             # e.g., "raw telemetry, pre-calibration"
  embargo_until: __YYYY-MM-DD__             # [if_applicable] Required if state=embargo.
                                             # ISO 8601 date after which release is permitted.
```

Datacard template

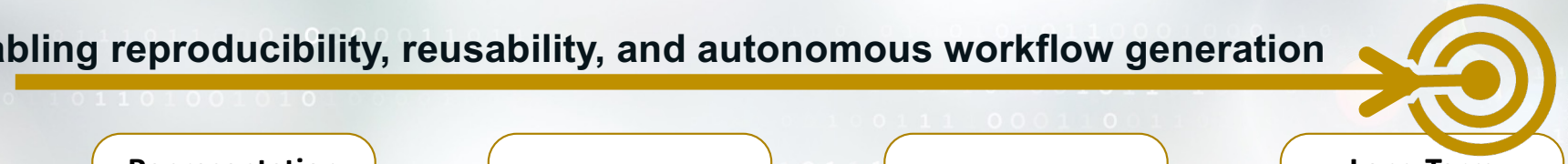
Documenting the qualified references for the dataset:

```yaml
related_resources:
datasets:
    - name: "SNS Raw Beam Position Monitor Telemetry 2023"
        identifier:
            type: ark
            value: ark:/12345/sns_bpm_raw_2023
        relationship: is_derived_from # dataset was produced from
    - name: "SNS BPM Calibration Reference Dataset"
        identifier:
            type: doi
            value: 10.25982/98765.4321
        relationship: is_based_on # calibration parameters sourced from
software:
    - name: "SNS BPM Calibration Pipeline"
        version: "2.1.0"
        identifier:
            type: url
            value: https://github.com/ornl-sns/bpm-calibration
        relationship: used_to_create # pipeline that generated dataset
```

```yaml
    - name: "PyBPM Analysis Toolkit"
        version: "1.4"
        identifier:
            type: doi
            value: 10.5281/zenodo.7891234
        relationship: used_to_analyze # used downstream for analysis
ai_models:
    - name: "BPM Anomaly Detector v1"
        version: "1.0"
        date_accessed: "2024-11-15"
        identifier:
            type: url
            value: https://huggingface.co/ornl-sns/bpm-anomaly-v1
        relationship: trained_on # this model was trained on this dataset
```

Datacard status and vision

Enabling reproducibility, reusability, and autonomous workflow generation



Current Step: Structured Dataset Documentation
Create structured metadata and datacards linking datasets to related resources.

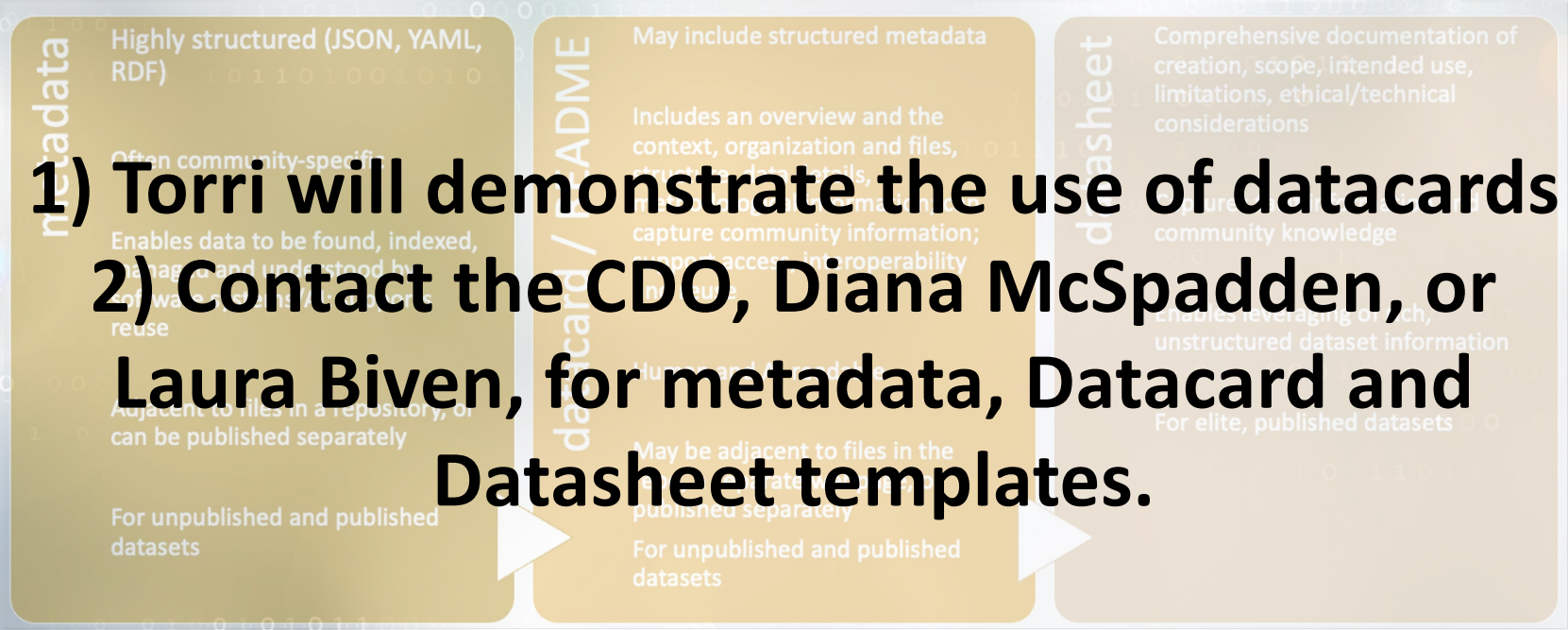
Representation Research Question
"What is the optimal way to represent dataset, software, and workflow relationships?"

Automation Research Question
"Can we automatically generate the information?"

Toward Autonomous Knowledge Graphs
Autonomously building the web of research products.

Long-Term Vision: Generative Workflows
"Can we enable generative development and reuse of scientific workflows?"

Metadata and Documentation



1) Torri will demonstrate the use of datacards

2) Contact the CDO, Diana McSpadden, or Laura Biven, for metadata, Datacard and Datasheet templates.

Jefferson Lab

Chief Data Office at JLab

JLAB Chief Data Office

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

Laura Biven
Chief Data Officer
orcid.org/0000-0002-5755-8449

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
https://orcid.org/0000-0002-8520-1631

My interests:

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

What we do (so far)

Cross-Lab Initiatives

Genesis Mission

U.S. DEPARTMENT OF ENERGY DOE DATA DAY (D3)

DOE Data Curation Working Group

JLab Guidance and Processes

JLab has developed a DMSP template to help you meet DMSP requirements.
https://sci-software.jlab.org/

Templates for Data Management and Sharing Plans (Nov 2025):

JLab DMSP Template and Guidance (PDF)
JLab DMSP Template and Guidance (TeX)

Data Stewards

- Leadership roles in Genesis ModCon
- Collaborators on active research projects for data management R&D
- Governance and data management support to research activities

Community Trends

FAIR DIGITAL OBJECTS FORUM

Alliance for Data Science and AI

RDA RESEARCH DATA ALLIANCE

- Frequent invited presentations
- Board activities

Kudos

Casey Morean
Physics AI/ML Data Scientist
orcid.org/0000-0001-5588-4841
My interests:
- ***Software supporting metadata capture and automation in science***
- ***AI infrastructure***
- ***Physics: EMC effect, SRCs, GPDs***

Anil Panta
Physics Data Scientist
orcid.org/0000-0001-6385-7712
My interests:
- ***Distributed Data Management***
- ***Document Management***
- ***Charm physics and CP violation***

Nataliia Matsiuk
Scientific Technical & Program Support
orcid.org/0000-0003-0306-5152
My interests:
- ***Software development***
- ***AI/ML***
- ***Painting, powerlifting, and planting trees***

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!
- As we highlight common code paths, the STRIDE team is available to set up shared infrastructure for the lab. Contact Brad Sawatzky brads@jlab.org or stride@jlab.org

STRIDE
Software Tools for Research, Infrastructure, Data, and Experiments

FAIR4…

FAIR for Research Software https://www.rd-alliance.org/groups/fair-research-software-fair4rs-wg/activity/

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

Jefferson Lab

Thank you

data.jlab.org

JLabDataManagement@jlab.org

U.S. Department of ENERGY

Backups

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***

Where is this going? (CDO's predictions)

- Increasing expectations for {data, code, documentation…} sharing
  - More data, more immediate sharing, more context,
  - Publications, data, code, and AI models are fully integrated, FAIR and AI-ready.
  - ***Build World-Class Scientific Datasets:*** *"*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." - https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/

- Tensions between protecting data and openness
- Enhanced interplay between humans and AI
  - Machines become better at understanding theory, not just data.
  - Increasing need for validation and resiliency for AI in the scientific process
- Emergence of global frameworks and standards for data [,workflows] and metadata
  - Shameless plug for FDO Conference: https://fairdo.org/fdo-conference-2026-registration/
  - Shameless plug for Research Data Alliance: https://www.rd-alliance.org/, https://zenodo.org/communities/rda/records

Data Management Planning

As part of research, experiment and collaboration planning

Jefferson Lab

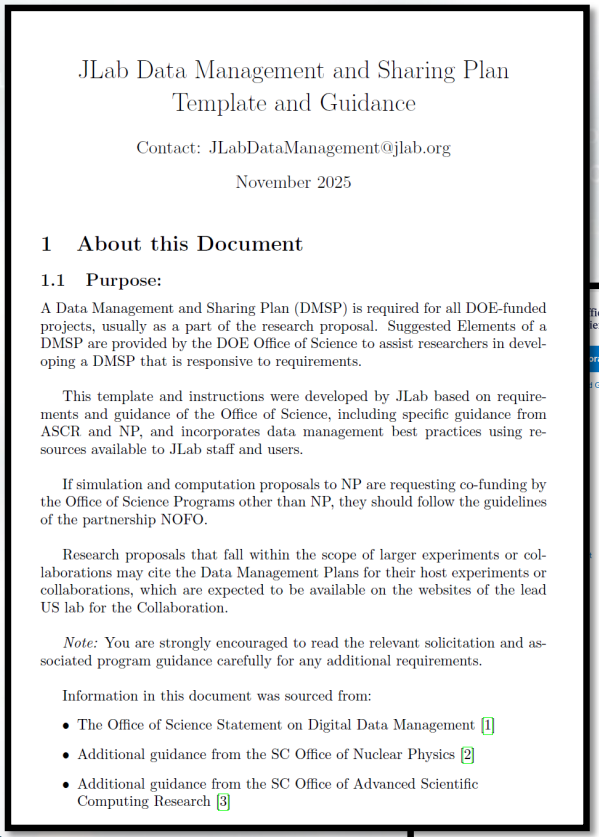
DOE and other federal funding agencies require **Data Management and Sharing Plans** as part of research proposals.

JLab has developed a DMSP template to help you meet DMSP requirements.

https://sci-software.jlab.org/

Templates for Data Management and Sharing Plans (Nov 2025):

JLab DMSP Template and Guidance (PDF)
JLab DMSP Template and Guidance (TeX)



JLab Data Management and Sharing Plan Template and Guidance

Contact: JLabDataManagement@jlab.org

November 2025

1 About this Document

1.1 Purpose:

A Data Management and Sharing Plan (DMSP) is required for all DOE-funded projects, usually as a part of the research proposal. Suggested Elements of a DMSP are provided by the DOE Office of Science to assist researchers in developing a DMSP that is responsive to requirements.

This template and instructions were developed by JLab based on requirements and guidance of the Office of Science, including specific guidance from ASCR and NP, and incorporates data management best practices using resources available to JLab staff and users.

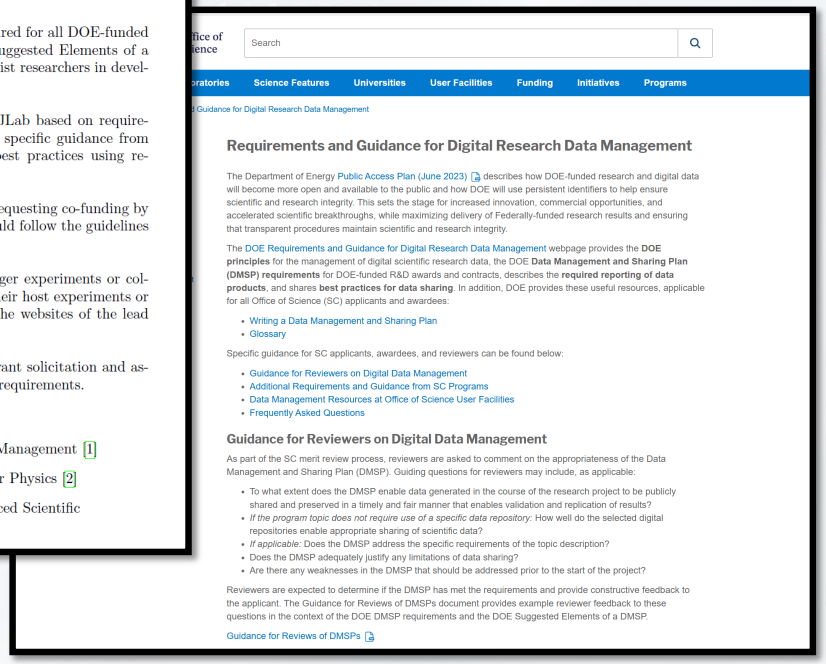
If simulation and computation proposals to NP are requesting co-funding by the Office of Science Programs other than NP, they should follow the guidelines of the partnership NOFO.

Research proposals that fall within the scope of larger experiments or collaborations may cite the Data Management Plans for their host experiments or collaborations, which are expected to be available on the websites of the lead US lab for the Collaboration.

Note: You are strongly encouraged to read the relevant solicitation and associated program guidance carefully for any additional requirements.

Information in this document was sourced from:

- The Office of Science Statement on Digital Data Management [1]
- Additional guidance from the SC Office of Nuclear Physics [2]
- Additional guidance from the SC Office of Advanced Scientific Computing Research [3]



Search

Science Features | Universities | User Facilities | Funding | Initiatives | Programs

Guidance for Digital Research Data Management

Requirements and Guidance for Digital Research Data Management

The Department of Energy Public Access Plan (June 2023) describes how DOE-funded research and digital data will become more open and available to the public and how DOE will use persistent identifiers to help ensure scientific and research integrity. This sets the stage for increased innovation, commercial opportunities, and accelerated scientific breakthroughs, while maximizing delivery of Federally-funded research results and ensuring that transparent procedures maintain scientific and research integrity.

The DOE Requirements and Guidance for Digital Research Data Management webpage provides the **DOE principles** for the management of digital scientific research data, the DOE **Data Management and Sharing Plan (DMSP) requirements** for DOE-funded R&D awards and contracts, describes the **required reporting of data products**, and shares **best practices for data sharing**. In addition, DOE provides these useful resources, applicable for all Office of Science (SC) applicants and awardees:

- Writing a Data Management and Sharing Plan
- Glossary

Specific guidance for SC applicants, awardees, and reviewers can be found below:

- Guidance for Reviewers on Digital Data Management
- Additional Requirements and Guidance from SC Programs
- Data Management Resources at Office of Science User Facilities
- Frequently Asked Questions

Guidance for Reviewers on Digital Data Management

As part of the SC merit review process, reviewers are asked to comment on the appropriateness of the Data Management and Sharing Plan (DMSP). Guiding questions for reviewers may include, as applicable:

- To what extent does the DMSP enable data generated in the course of the research project to be publicly shared and preserved in a timely and fair manner that enables validation and replication of results?
- *If the program topic does not require use of a specific data repository:* How well do the selected digital repositories enable appropriate sharing of scientific data?
- *If applicable:* Does the DMSP address the specific requirements of the topic description?
- Does the DMSP adequately justify any limitations of data sharing?
- Are there any weaknesses in the DMSP that should be addressed prior to the start of the project?

Reviewers are expected to determine if the DMSP has met the requirements and provide constructive feedback to the applicant. The Guidance for Reviews of DMSPs document provides example reviewer feedback to these questions in the context of the DOE DMSP requirements and the DOE Suggested Elements of a DMSP.

Guidance for Reviews of DMSPs

https://science.osti.gov/Funding-Opportunities/Digital-Data-Management

Persistent, Unique Identifiers and a Web of Scientific Artifacts

Jefferson Lab

Digital persistent identifier (DPI or digital PID) – A digital identifier that is globally unique, persistent, machine resolvable and processable, and has an associated metadata schema.*

- 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

Satisfies DOE Requirement

- 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.**

* DOE Public Access Plan 2023 https://doi.org/10.11578/2023DOEPublicAccessPlan

FAIR for Data

https://www.gofair.foundation/interpretation

Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., Honeyman, T., Struck, A., Lee, A., Loewe, A., van Werkhoven, B., Jones, C., Garijo, D., Plomp, E., Genova, F., … RDA FAIR4RS WG. (2022). FAIR Principles for Research Software (FAIR4RS Principles) (1.0). Zenodo. https://doi.org/10.15497/RDA00068

Barker, M., Chue Hong, N. P., Katz, D. S., Lamprecht, A.-L., Martinez, C., Psomopoulos, F., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., & Honeyman, T. (2022). FAIR principles for research software (FAIR4RS principles). Scientific Data, 9, 622. https://doi.org/10.1038/s41597-022-01710-x

FAIR Guiding Principles
Findable: The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process. E.g. **F4: (Meta)data are registered or indexed in a searchable resource**
Accessible: Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation. E.g. **A1: (Meta)data are retrievable by their Identifier using a standardized communications protocol.**
Interoperable: The data usually needs to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing. E.g., **I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.**
Reusable: The ultimate foal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined.

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

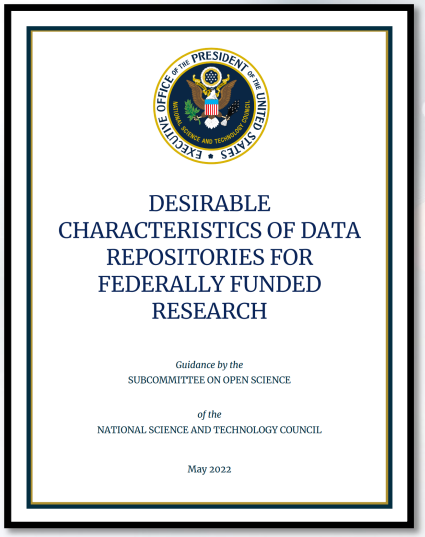
Brewer, W., Widener, P., Anantharaj, V., Wang, F., Beck, T., Shankar, A., & Oral, S. (2025, September). Data Readiness for Scientific AI at Scale. In Workshop Proceedings of the 54th International Conference on Parallel Processing (pp. 18-24). DOI:10.1145/3750720.3757282

Metadata and Repositories

- Repositories should have requirements for metadata.
- Domain-specific repositories will have more specific requirements and curation support.
- Metadata are absolutely essential and typically insufficient for reuse and interoperability

Notes:

- Submitting to a community repository is best practice.
- Generalist repositories are acceptable.
- Posting to a website is **NOT** best practice.



Metadata: The repository ensures datasets are accompanied by metadata to enable discovery, reuse, and citation of datasets, using schema that are appropriate to, and ideally widely used across, the communities that the repository serves.

https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf

What is Research Software?

[1] Hettrick, S., Antonioletti, M., Carr, L., Chue Hong, N., Crouch, S., De Roure, D., Emsley, I., Goble, C., Hay, A., Inupakutika, D., Jackson, M., Nenadic, A., Parkinson, T., Parsons, M. I., Pawlik, A., Peru, G., Proeme, A., Robinson, J., & Sufi, S. (2014). UK Research Software Survey 2014 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14809

[2] Software and Source Codes College. https://www.ouvrirlascience.fr/research-software-as-a-pillar-of-open-science/.

U. Nangia and D. S. Katz, "Understanding Software in Research: Initial Results from Examining Nature and a Call for Collaboration," 2017 IEEE 13th International Conference on e-Science (e-Science), Auckland, New Zealand, 2017, pp. 486-487, doi: 10.1109/eScience.2017.78.

What is Research Software?

- Software that generates, processes, or analyzes results to enable publishing scientific research [1]
- Can exist throughout the research lifecycle (before, during, alongside, after publication) [2]

Software is 1) a product of and 2) a research process.

FAIR vs. FAIR4RS

[1] Lamprecht, A. L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., ... & Capella-Gutierrez, S. (2020). Towards FAIR principles for research software. Data Science, 3(1), 37-59. https://doi.org/10.3233/DS-190026.
[2] Allen, R., & Hartland, D. (2018). FAIR in practice-Jisc report on the findable accessible interoperable and reuseable data principles. https://doi.org/10.5281/zenodo.1245567.

"interoperability has been found to be 'the most challenging of the four FAIR principles. This, in part, is due to interoperability not being well understood' [2]. In contrast to the rather static nature of data, research software are live digital objects that interact at different levels with other objects, e.g., other software, managed data, execution environments; and either directly and/or indirectly, as scripts or as part of a workflow (see Fig. 1). The interoperability principles are therefore even more challenging to apply to software, **some are not directly applicable, others need to be rephrased and even new principles need to be defined to appropriately address the dynamic nature of software.**" [1]

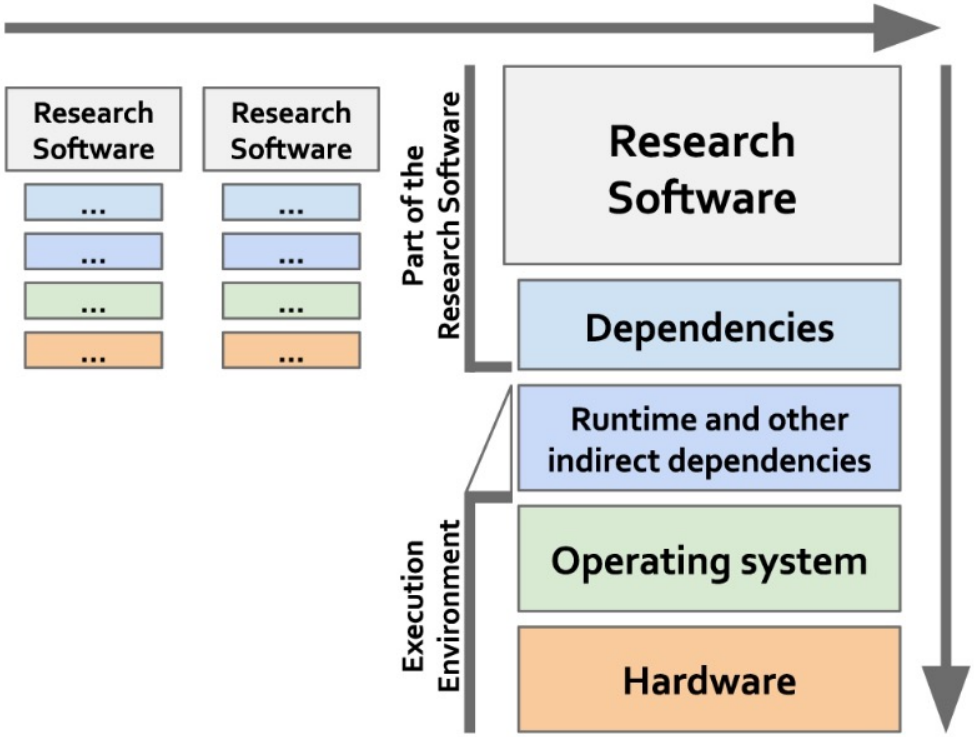


Fig. 1. Interoperability for research software can be understood in two dimensions: as part of workflows (horizontal dimension) and as stack of digital objects that need to work together at compilation and execution times (vertical dimension). Importantly, workflows do not need to use the same physical hardware or the same operating system, as long as there are agreed mechanisms for software to interoperate with one another.

Data Analysis and Preservation; and Provenance

- Build software into containers and share those versioned containers in GitLab
  - Documentation on GitLab usage at Jefferson Lab
  - Documentation on Container usage at Jefferson Lab
- Take the time to tag the analysis with metadata for analysis discovery.
- Separation of Data (available through Rucio), Configuration (YAML, JSON format), and Software (containers)
- RO-Crate designed for packaging digital objects (data, software, instruments, licenses, workflow executions) along with their metadata, thereby facilitating the capture of dependencies and context.

Peroni S, Soiland-Reyes S, Sefton P, et al. Packaging research artefacts with RO-Crate. Data Science. 2022;5(2):97-138. doi:10.3233/DS-210053

Leo S, Crusoe MR, Rodríguez-Navas L, Sirvent R, Kanitz A, De Geest P, et al. (2024) Recording provenance of workflow runs with RO-Crate. PLoS ONE 19(9): e0309210. https://doi.org/10.1371/journal.pone.0309210

Soiland-Reyes, S., Sefton, P., Leo, S., Castro, L. J., Weiland, C., & Van de Sompel, H. (2025). Practical webby FDOs With RO-Crate and FAIR Signposting: Experiences and Lessons Learned. *Open Conference Proceedings*, *5*. https://doi.org/10.52825/ocp.v5i.1273