ModCon Metadata and Datacards

RHIC DAP Roundtable

06/18/2026

Diana McSpadden

Data Scientist – Data Steward, JLAB

Jefferson Lab

U.S. Department of ENERGY

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

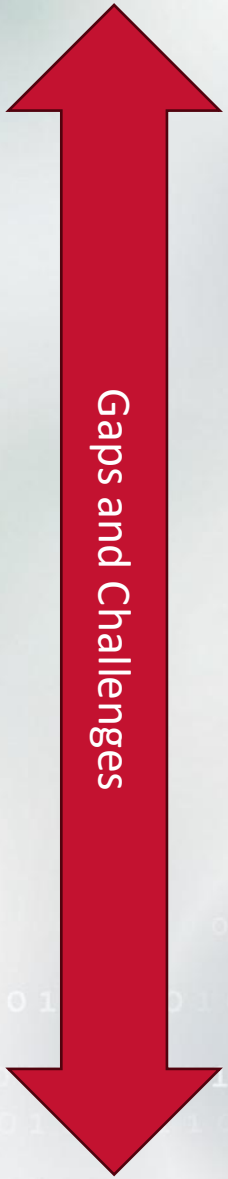
My interests:

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

1. Requirements in Data Management and Sharing

2. Genesis Mission through the data lens

3. FAIR and AI-Ready Data and the ModCon Datacard



Data Management and Sharing Requirements

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***

Gaps, Challenges, and Opportunities

1. Each community needs to establish norms for **“data underlying a publication”** and revisit these as technologies and expectations change.

2. Each community needs to establish norms for **data sharing timelines** and revisit these as technologies and expectations change.

3. **Trusted Repositories** that steward data long-term are needed.

Opportunity: Build momentum for data documentation, data sharing, and data stewardship to advance dataset use, re-use, and workflow/pipeline automation.

LLNL-PRES-2016998

Genesis Mission

About the Genesis Mission

The **Genesis Mission** is a historic national effort to catalyze new industries, create high-skill jobs, and usher a new golden era of American discovery through artificial intelligence innovation.

The Genesis Mission's **American Science and Security Platform** will connect the world's best supercomputers, AI systems, and next-generation quantum computers with the most exquisite scientific instruments in the nation, and its intelligence layer will be trained with the singular scientific datasets and expertise housed in the National Laboratories.

Once complete, it will be the world's most complex and powerful scientific instrument ever built.

FAIR, AI-Ready Data, and the ModCon Datacard

Diana's 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

Data Context for machine actionability

Data useful, *when*
Data in context

- What is it?
- How was it created?
- Where is it stored?
- How persistent is it?
- Who is allowed to access it?
- How to reuse it?
- How to interpret it?
- With which operation can we process it?

FAIR Metadata + LLM-ready Narrative

Courtesy of D. Koureas, 2024 FDO Forum https://fairdo.org/fdof-summit-2024/
https://eosc.eu/wp-content/uploads/2024/10/Koureas_FAIR-Digital-Objects.pdf

Context: 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

data card / 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?

Six intended capabilities explicitly indicate what the data card intends to support:

Intended Capability	Goal
Discoverability	Required fields are the minimum metadata for a dataset; enough metadata to find and identify the dataset Optional fields include the workflow state of the dataset.
Accessibility	Required fields include details for how humans and agents can access a dataset and any access restrictions. Optional fields include details of the approved environments, policy, and access end points.
Interoperability	Required fields describe the characteristics such as the dataset structure, its modalities, and generation. Optional fields include source data sets and simulation details
Reusability	Required fields include data quality elements. Optional fields include the license(s) and stewardship details.
Governed Use	Required fields include a description of current use, export control, privacy status, compliance information Optional fields include intended, permitted, and prohibited use of the data.
AI Usability	Required fields specify whether training, inference, and evaluation are permitted, as well as any AI-related restrictions, biases, and safety considerations.

Data card template

The template annotates every field: [capability].[required | if_applicable]

```
        value: __VALUE__
# --- Identification -------------------------------------------
identification:                          # [discoverability_required] Key metadata fields that uniquely identify
this dataset.
  name: "${DATASET_NAME}"                # [discoverability_required] Single human-readable name for this dataset.
                                         # Use the same name in the data card filename.
                                         # If this data card covers a collection, provide the collection name.
  project: "${PROJECT_NAME}"             # [discoverability_required] Genesis project or sub-project this dataset
  belongs to.
                                         # e.g., genesis | genesis-fusion | genesis-lightsource
  version: "1.0"                         # [discoverability_required] Dataset version using semantic versioning:
  MAJOR.MINOR.PATCH
```

```
ai_usability:
  # --- AI / ML Usage ------------------------------------------
  # [ai_ready] Describes whether and how this dataset may be used
  # in AI/ML workflows. Be explicit — these fields are read by
  # automated pipeline tooling.
  ai_usage:
    training_use_allowed: __ALLOWED__          # [ai_usability_required] "Yes" | "No" | "Conditional"
    inference_use_allowed: __ALLOWED__         # [ai_usability_required] "Yes" | "No" | "Conditional"
    evaluation_use_allowed: __ALLOWED__        # [ai_usability_required] "Yes" | "No" | "Conditional"
    restrictions: __DESCRIPTION__              # [ai_usability_required] e.g., "Not for clinical decision-making" | "None"
    bias_risks: __DESCRIPTION__                # [ai_usability_required] e.g., "Overrepresents samples from facility X" |
    "None"
    safety_considerations: __DESC__            # [ai_usability_required] e.g., "Outputs may be export-controlled" | "None"
    human_review_required: "__Yes|No__"        # [ai_usability_required] "Yes" | "No"
```

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
    - 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
```

Data card template and schema location

Preparing public release

American Science Cloud GitLab: https://gitlab.com/amsc2/modcon/dbs/data-cards

Repo for Genesis data card template and schema: documentation for scientific datasets that supports discovery, access, interoperability, reusability, governed use, and AI use.

Contains resources for aligning data cards with the Genesis ecosystem:

- Templates
- Examples
- LinkML schema

The LinkML schema for the data card:

- defines the structure and constraints for the data cards through classes:, enums:, and slots:; required fields, data types, and validation rules
- Valid completed data cards conform to this schema.
- Rich documentation for each field

Datacard status and possible vision:

Phase 1: Now

Datacard Template

Corpus of Datacards produced ***post hoc*** from humans, papers, etc.

Test the utility of data card content with use-case based metrics

Assess the effectiveness of post hoc datacard creation methods

Datacard status and possible vision:

Phase 1: Now	Phase 2: Lifecycle Integration of xCards
Datacard Template	xCard Template with customization
Corpus of Datacards produced ***post hoc*** from humans, papers, etc.	xCards created as part of the lifecycle
Test the utility of data card content with use-case based metrics Assess the effectiveness of post hoc datacard creation methods	xCards enable autonomous search, access, reuse of digital objects.

Gaps and Challenges

1. Each community needs to establish norms for **“data underlying a publication”** and revisit these as technologies and expectations change.

2. Each community needs to establish norms for **data sharing timelines** and revisit these as technologies and expectations change.

3. **Repositories** that steward data long-term are needed. // Take on long-term stewardship for digital assets.

4. Enormous pressure to provide metadata and context for data and code to make them FAIR and AI-Ready.

5. Develop **context for digital object throughout the data lifecycle**. Humans should write ideas once.

Datacard status and possible vision:

Phase 1: Now	Phase 2: Lifecycle Integration of xCards	Phase 3a: "Web of Science"
Datacard Template	xCard Template with customization	PIDs + xCards + controlled relationships create a web of digital assets
Corpus of Datacards produced ***post hoc*** from humans, papers, etc.	xCards created as part of the lifecycle	Web of digital objects and interoperability enables more efficient computation for AI training and workflows
Test the utility of data card content with use-case based metrics Assess the effectiveness of post hoc datacard creation methods	xCards enable autonomous search, access, reuse of digital objects.	

Datacard status and possible vision:

Phase 1: Now

Datacard Template

Corpus of Datacards produced ***post hoc*** from humans, papers, etc.

Test the utility of data card content with use-case based metrics

Assess the effectiveness of post hoc datacard creation methods

Phase 2:

xCard Template with customization

xCards created as part of the lifecycle

xCards enable autonomous search, access, reuse of digital objects.

Phase 3a: "Web of Science"

PIDs + xCards + controlled relationships create a web of digital assets

Web of digital objects and interoperability enables more efficient computation for AI training and workflows

Phase 3b: Generative Workflows

PIDs + xCards + provenance enable reproducible workflows

Workflows themselves are training data for GenAI capabilities

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

Data Governance, Sharing, and Reuse

- Outline data governance rules at the outset.
- Instead of one-off data sharing with colleagues, consider publishing data to a repository.
- For highly-reuseable data, consider developing a datasheet or data publication.
- Do cite data in your publications.

Data Analysis and Preservation; and Provenance

- Build software into containers and share those versioned containers in GitLab. Follow FAIR4RS guidelines.
- Include provenance capture as part of data management planning at the outset.
- Build xCards as you go.
- Test, test, test for reproducibility, machine actionability, AI readiness,…

Jefferson Lab

Thank you

data.jlab.org

U.S. Department of ENERGY

Some Terminology

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/)

Definitions in the OPEN Government Data Act

Data: recorded information, regardless of form or the media on which the data is recorded

Data Asset: a collection of data elements or data sets that may be grouped together

Machine-Readable Data: data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost

Public Data Asset: a data asset, or part thereof, maintained by the Federal Government that has been, or may be, released to the public, including any data asset, or part thereof, subject to disclosure under FOIA

Open Government Data Asset: a public data asset that is (A) machine-readable; (B) available (or could be made available) in an open format; (C) not encumbered by restrictions, other than intellectual property rights … that would impede the use or reuse of such asset; and (D) based on an underlying open standard that is maintained by a standards organization

Metadata: structural or descriptive information about data such as content, format, source, rights, accuracy, provenance, frequency, periodicity, granularity, publisher or responsible party, contact information, method of collection, and other descriptions

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/

The Genesis Mission:
Prioritizing American Science and Technology Leadership

"The Genesis Mission …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 Genesis Mission will create "…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

Genesis Mission

National Initiative for Science, National Security, and Innovation

Mission Goal: *Use artificial intelligence to dramatically accelerate American scientific discovery and* ***strengthen*** *U.S. economic and national security leadership.*

The Challenge

A global race amidst a revolution in computing

- Modern AI is bringing a completely new way to conduct science
- Nations compete to transform the speed and quality of scientific achievement
- We must deliver world-changing scientific AI capabilities targeted at the nation's most critical science and security needs

The Solution

A portfolio of National Science and Technology Challenges delivered by an integrated AI-driven platform

- National Challenges define where the US must lead in science, energy, and national security
- The American Science and Security Platform uses AI to combine federal data, computing, laboratories, and facilities into an exquisite, unified discovery instrument
- Automates and accelerates research, experimentation, and engineering
- Keeps **humans in control** while removing workflow bottlenecks

What This Delivers

- **Faster results:** 10–100× acceleration in priority science and engineering domains
- **Stronger security:** AI-enabled solutions and threat mitigation in high-consequence missions
- **Better returns:** Multiplies impact of existing federal R&D investments
- **U.S. leadership:** Technological superiority across the globe

Close industry partnerships, delivered by the **partnerships team**, will enable rapid platform development and ensure meaningful application impact

U.S. DEPARTMENT of ENERGY

LLNL-PRES-2016998

Genesis Mission

The Genesis Mission is organized around two concepts

The National Science and Technology Challenges

The **National Challenges** are high-impact technical problems aligned to urgent national priorities – where AI can dramatically accelerate progress. They demand AI model, data, computing, and automation capabilities unified in a single solution platform. Expert teams develop innovative solutions while results strengthen and extend the platform itself to tackle ever more ambitious challenges.

Delivered by the National Challenges team through the platform.

The American Science and Security Platform

The **platform** supports AI-driven experimentation, analysis, discovery, design, and manufacturing. It unifies access to AI, computing resources, scientific data, and automated facilities, allowing application of unprecedented capabilities. It delivers scientific self-improving feedback that will accelerate R&D in the energy dominance, discovery science, and national security pillars.

Delivered by the infrastructure, data, and models teams.

Close industry partnerships, delivered by the **partnerships team**, will enable rapid platform development and ensure meaningful pillar impact

Genesis Mission

The mission technical efforts are organized into five teams

National Challenges

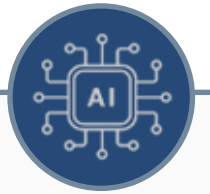
Engages with projects addressing high-impact national challenges—such as energy security and advanced materials—that are accelerated by the platform. Serves as a living portfolio that evolves as goals and technologies change.

Infrastructure

Provides hardware and software foundations to support large-scale AI training and inference, data management, and scalable agentic workflows. Ensures security of models, data, and systems and interoperability across DOE facilities and production agencies.

Data

Organizes, curates, and prepares scientific and engineering data from experiments, simulations, and observations for AI use. Ensures high-quality, reliable data to accelerate discovery and problem-solving.



Models

Develops an AI ecosystem of agents and models that combines frontier industry capability with DOE-specific expertise to discover and implement novel solutions to the nation's open science and technology challenges.

Partnerships

Builds collaborations across government, industry, and academia to advance AI innovation. Leverages shared expertise and resources to deliver real-world impact at national scale.