



Data Management and Stewardship

CDAO Government

September 22, 2026

Laura Biven

Chief Data Officer, Jefferson Lab

 Jefferson Lab

U.S. Department of
ENERGY

Laura Biven, Chief Data Officer

biven@jlab.org; <https://data.jlab.org/>; orcid.org/0000-0002-5755-8449



Advanced Scientific Computing
Research (ASCR)

DOE Requirements and
Guidance for Digital
Research Data
Management

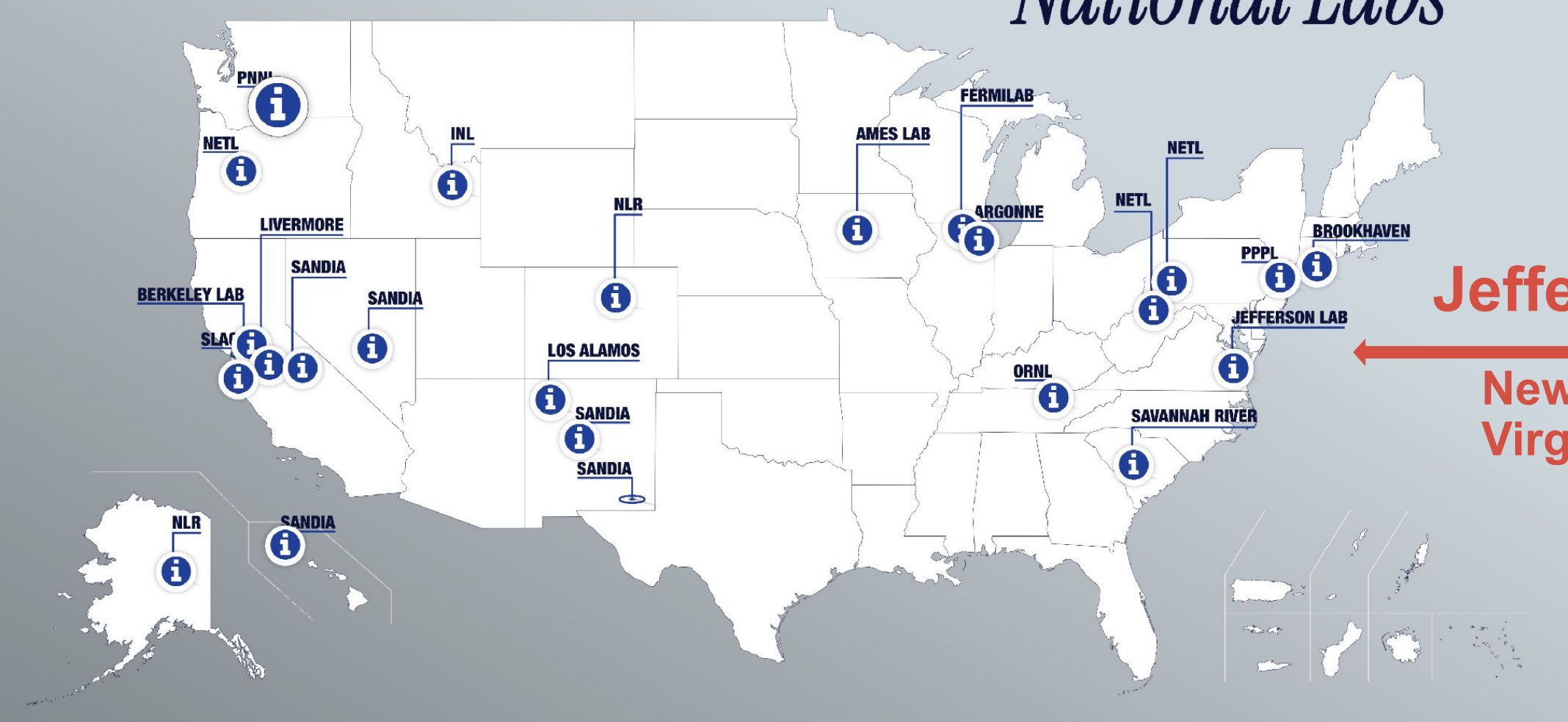


My interests:

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



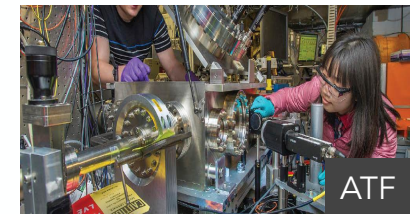
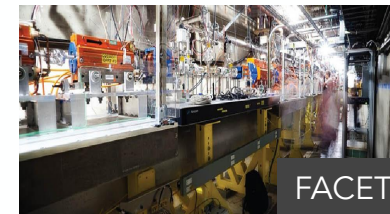
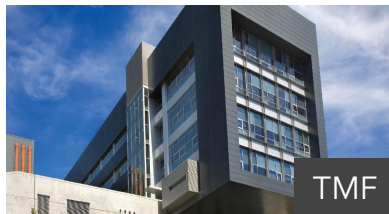
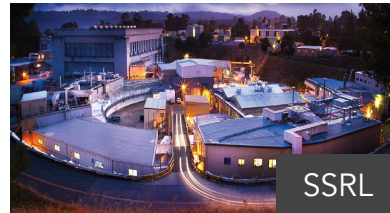
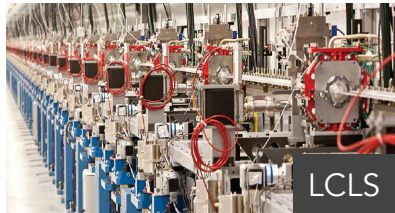
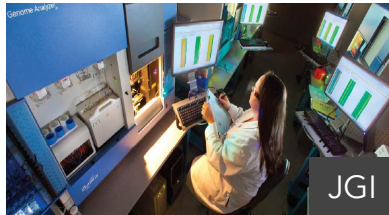
U.S. DEPARTMENT of ENERGY *National Labs*



Jefferson Lab
Newport News,
Virginia

The Office of Science User Facilities

FY 2025
28 scientific
user facilities
>39,500 users

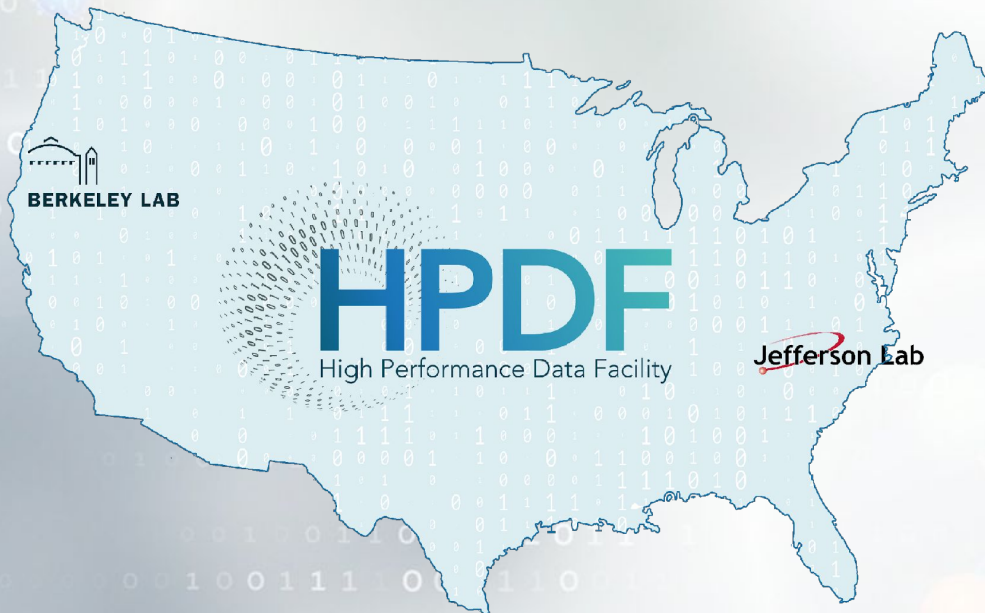
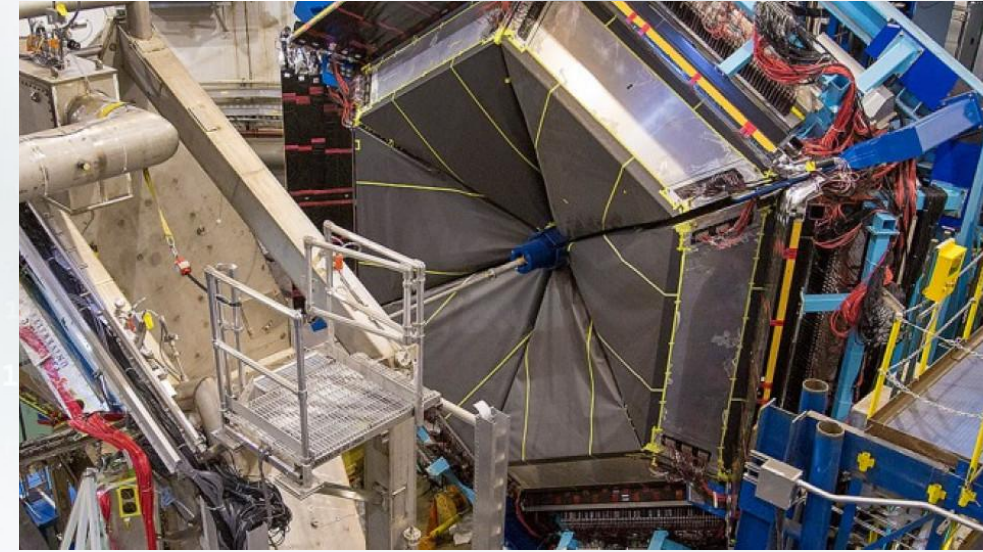


U.S. DEPARTMENT
of **ENERGY** | Office of
Science

Jefferson Lab

Jefferson Lab

Thomas Jefferson National Accelerator Facility (Jefferson Lab) is a U.S. Department of Energy Office of Science national laboratory. Scientists worldwide utilize the lab's unique particle accelerator, known as the **Continuous Electron Beam Accelerator Facility (CEBAF)**, to probe the most basic building blocks of matter.



DOE Selected Thomas Jefferson National Accelerator Facility as the lead for its new **High Performance Data Facility (HPDF)** Hub. Jefferson Lab will partner with DOE's Lawrence Berkeley National Laboratory to form a joint project team led by Jefferson Lab. The HPDF will be a computing and data infrastructure resource that will provide transformational capabilities for data analysis, networking and storage for the nation's research enterprise.

<https://www.jlab.org/>

<https://science.osti.gov/np/Facilities/User-Facilities/CEBAF>

<https://www.hpdp.science/>

- 
- 
- 
1. Motivation
 2. AI Ready Data
 3. Data Cards
 4. Next Steps and Parting Thoughts

Motivation

Genesis Mission - National Science & Technology Challenges

Helping Americans Live Longer, Healthier Lives

- Finding the Root Causes of Chronic Disease
- Accelerating Drug Discovery and Clinical Translation
- Unlocking Cures for Pediatric Cancer
- Delivering Better Health Outcomes for Veterans

Building American Industrial Strength

- Reenvisioning Advanced Manufacturing and Industrial Productivity
- Scaling Biology for American Industrial Leadership
- Securing America's Critical Minerals Supply
- Recentering Microelectronics in America
- Securing U.S. Leadership in Data Centers

Delivering Reliable Infrastructure and Energy Affordability

- Reimagining the Lifecycle of American Infrastructure
- Delivering Nuclear Energy That is Faster, Safer, Cheaper
- Accelerating Delivery of Fusion Energy
- Transforming Nuclear Cleanup and Restoration
- Predicting U.S. Water for Energy
- Scaling the Grid to Power the American Economy
- Unleashing Subsurface Strategic Energy Assets

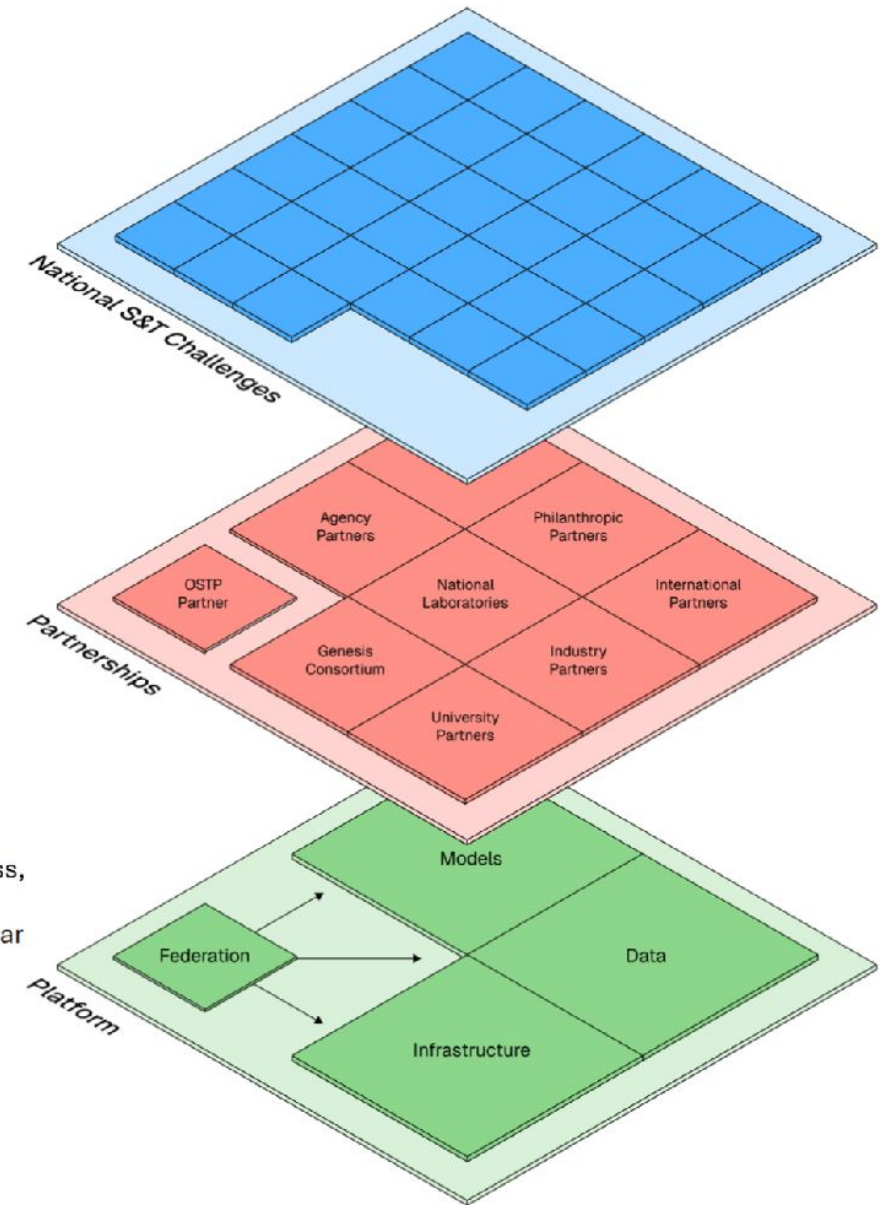
<https://www.energy.gov/documents/genesis-mission-national-science-technology-challenges>

Extending the Frontiers of American Discovery

- Discovering Quantum Algorithms with AI
- Realizing Quantum Systems for Discovery and Use
- Achieving AI-Driven Autonomous Laboratories
- Predicting Living Systems
- Designing Materials with Predictable Functionality
- Enhancing Particle Accelerators for Discovery
- Unifying Physics from Quarks to the Cosmos
- Mining Decades of Space Mission Data for Discovery

Protecting the Nation from Emerging Threats

- Early Detection and Attribution of Biological Threats
- Accelerating Design of Weapon Components and Systems
- Accelerating Materials Discovery, Production, and Qualification for Strategic Deterrence
- Accelerating Nuclear Threat Assessment, Preparedness, and Response
- Strengthening Deterrence Through Attribution of Nuclear Signatures
- Detecting, Locating, Identifying, and Characterizing Proliferation Threats
- Accelerating Experimental, Design, Production, and Nuclear Deterrence Outcomes
- Software Understanding for National Security
- Ensuring American Space Superiority



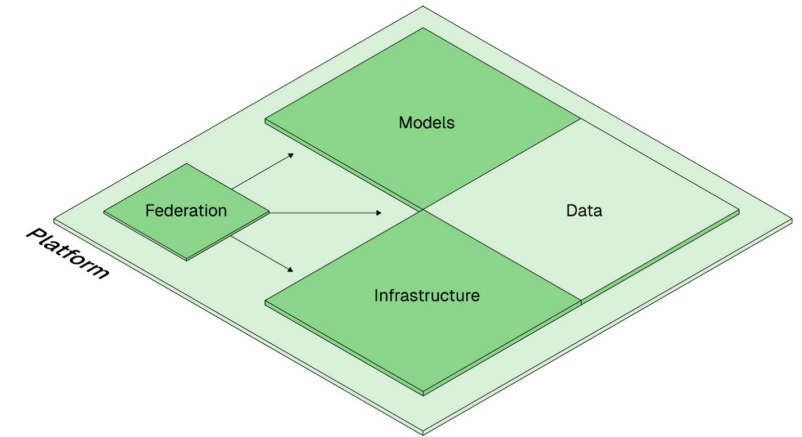
Genesis Mission – Data in the Platform

Federation

- The connections between platform data elements are critical
- Genesis Mission data are bigger than any single institution — data elements span universities, labs, agencies, and industry, each with its own priorities and constraints...

Types of Data (examples)

- Public vs Non-public
- Open data
- Unlimited Rights / Limited Rights
- Sensitive or Protected Data
 - Working Data (Non-public)
 - (CUI, Classified, Proprietary etc)
- Existing, Historical Data
- New Data
- Metadata
- Domain Data, e.g. Streaming / Automated Data



Data to Enable Partnerships & Managed Use

- Agreements
- Attribution
- Education & Outreach
- Governance
- Intellectual Property
- Lineage (Persistent Identifiers)
- Policies

Data to Drive Impact

- Fueling Models
- Benchmarks
- Results
- Validation

Data as Infrastructure

- Software and Services
- Agents
- Networks
- Storage & Curation
- Security
- 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 Genesis Mission will produce "the world’s largest and highest-quality scientific data sets to train the next generation of AI systems".

- Under Secretary for Science, Darío Gil
<https://www.energy.gov/science/articles/under-secretary-gils-letter-community>



AI-Ready Data

What Makes Data AI-Ready

Christine's

Working Definitions: AI Readiness for Data & AI Ready Data

AI Readiness for Data refers to a spectrum of how ready a given set of data is to be used by an AI process without need for further pre-processing or curation.

When data is ready to be used without further work, it can be described as *AI Ready Data*. It is the nature of technology to evolve and thus, AI Readiness is likely a near continuous process that must be undertaken and iterated for data to remain AI Ready.

Source: Upcoming CODATA Concept Paper

Courtesy of Christine Kirkpatrick, "El Dorado or Troy?
Demystifying FAIR and AI Ready Data" Data Readiness for
AI Workshop (DRAI) at ICPP'25

Element level:

- Data format
- Metadata completeness, accuracy
- Access...

Dataset level:

- Balance or representativeness (bias)
- Findability
- Ability to generate more data...

Data for Model Training

Data for Inference

MA Data for Agents

What Makes Data AI-Ready

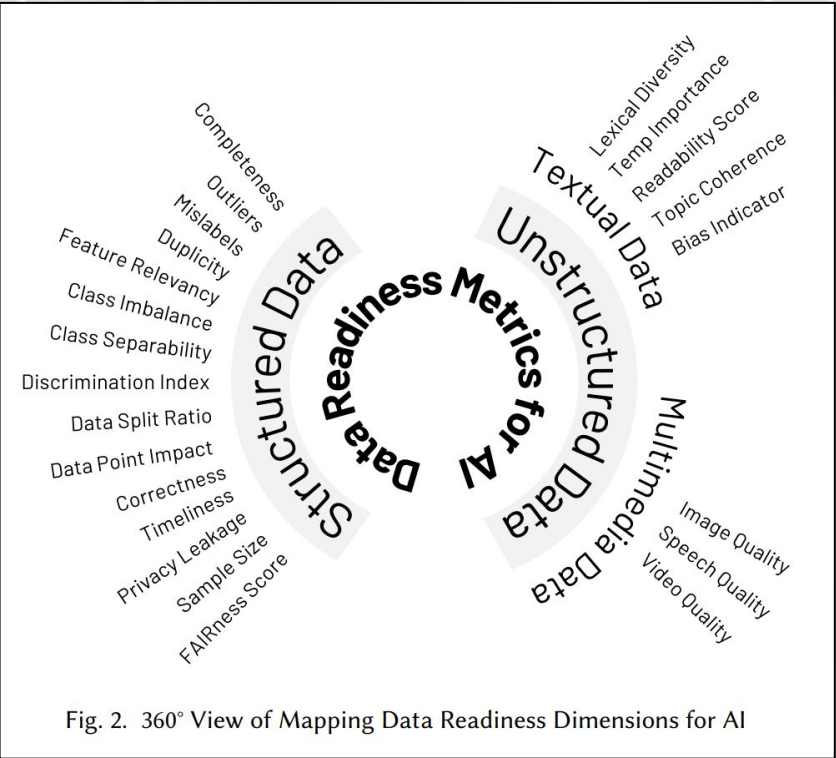


Table 2. 2D conceptual maturity matrix integrating Data Readiness Levels and Data Processing Stages (grey cells N/A)

Level	Ingest	Preprocess	Transform	Structure	Shard
1 - Raw	Initial raw acquisition				
2 - Cleaned	Validated ingestion into standard formats	Initial spatial/temporal alignment or regridding			
3 - Labeled	Enhanced metadata enrichment	Refined alignment; grids standardized	Initial normalization or anonymization; basic labels added		
4 - Feature-engineered	Optimized high-throughput ingestion	Alignment fully standardized	Normalization or anonymization finalized; comprehensive labeling	Domain-specific feature extraction completed	
5 - Fully AI-ready	Ingestion pipelines fully automated and performance-optimized	Alignment integrated and automated	Normalization / anonymization fully automated and audited	Feature extraction automated and validated	Data partitioned into train/test/val & sharded into binary formats for scalable ingestion

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

What Makes Data AI-Ready

Context for data is key to assessing and improving AI-readiness

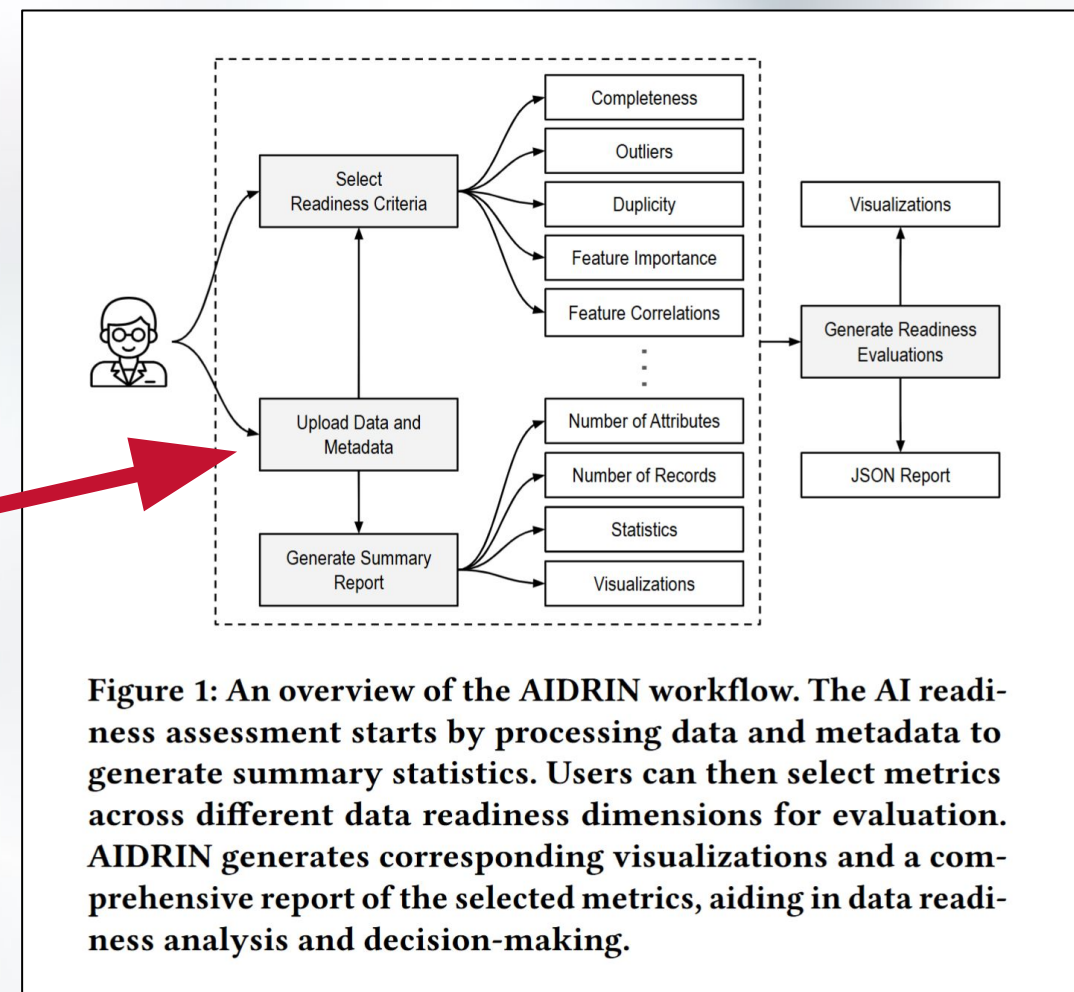


Figure 1: An overview of the AIDRIN workflow. The AI readiness assessment starts by processing data and metadata to generate summary statistics. Users can then select metrics across different data readiness dimensions for evaluation. AIDRIN generates corresponding visualizations and a comprehensive report of the selected metrics, aiding in data readiness analysis and decision-making.

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>

Data Cards

Data Cards

Challenge

- Critical know-how remains human-held
- Rate-limiters: Finding & assessing fit-for-purpose data
- AI and humans need explicit context:
 - What do they mean?
 - Reuse rules
 - Intended and actual use

Solution

- Rich and unambiguous context
- Machine-actionable data context *at scale*
- Relationships for a traversable web of scientific information

Data Cards for Standardized Metadata Across DOE-Aligned Data Initiatives: Toward Transparent, Interoperable, and Governed Dataset Documentation. DOI:

<https://doi.org/10.2172/3377514>

Genesis data card schema, template and supporting tools [Dataset]. Energy Data eXchange (EDX), Jefferson Lab. <https://doi.org/10.18141/3376151>.

Data Cards for Standardized Metadata Across DOE-Aligned Data Initiatives: Toward Transparent, Interoperable, and Governed Dataset Documentation

Document Version: 1.0

Date: July 2026



Prepared by:

Diana McSpadden, Thomas Jefferson National Accelerator Facility, Newport News, VA, USA

ORCID: <https://orcid.org/0000-0002-8520-1631> | ROR: <https://ror.org/02vwzrd76>

Olga Kuchar, Oak Ridge National Laboratory, Oak Ridge, TN, USA

ORCID: <https://orcid.org/0009-0000-2475-4710> | ROR: <https://ror.org/01qz5mb56>

Charlotte Neher, Sandia National Laboratories, Albuquerque, NM, USA

ORCID: <https://orcid.org/0009-0009-2169-0032> | ROR: <https://ror.org/01apwpt12>

Victor Walker, Idaho National Laboratory, Idaho Falls, ID, USA

ORCID: <https://orcid.org/0000-0001-5172-4410> | ROR: <https://ror.org/00ty2a548>

Jean Luca Bez, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

ORCID: <https://orcid.org/0000-0002-3915-1135> | ROR: <https://ror.org/02jbv0t02>

Laura Biven, Thomas Jefferson National Accelerator Facility, Newport News, VA, USA

ORCID: <https://orcid.org/0000-0002-5755-8449> | ROR: <https://ror.org/02vwzrd76>

Data Cards

Core questions a data card must answer:

- What is it?
- How was it created?
- Where is it stored?
- How persistent is it?
- Who can access it?
- How do I reuse/interpret it?
- With which operation can I process it?

FAIR Metadata

+

LLM-ready Narrative

+

Relationships

Data Cards for Standardized Metadata Across DOE-Aligned Data Initiatives: Toward Transparent, Interoperable, and Governed Dataset Documentation. DOI: <https://doi.org/10.2172/3377514>

Genesis data card schema, template and supporting tools [Dataset]. Energy Data eXchange (EDX), Jefferson Lab. <https://doi.org/10.18141/3376151>.

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

What Is the Data Card?

- It's a vehicle for collecting, organizing, validating, and presenting information needed to make a digital resource machine actionable and AI-Ready.
- Combines best practices of structured metadata block (YAML) and computable natural language (Markdown)
 - Machine-readable and human-readable,
 - For published and unpublished data.
- Not a replacement for repository-native metadata, legal review, or governance processes.
 - The data card is a common, extensible “perception layer” containing the context of the dataset.

Standards Alignment

- DCAT 3.0 [1]— interoperability baseline (title, description, identifier, creator, distribution...)
- DCAT-US [2] — U.S. federal catalog practice alignment
- OSTI / DataCite [3] — DOI issuance, relationship vocabulary
- ISO/IEC 11179 family [4] — metadata registry concepts (identification, classification, stewardship, lifecycle) — lightweight, not a full registry implementation
- Alexandria (NNSA) [5]— implementation reference for sensitivity/governance metadata

[1] W3C, Data Catalog Vocabulary (DCAT) – Version 3, W3C Recommendation, 2024

[2] General Services Administration, "DCAT-US Schema v1.1 (Project Open Data Metadata Schema)," resources.data.gov. [Online]. Available: <https://resources.data.gov/standards/catalog/dcat-us/>

[3] DataCite Metadata Working Group, *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs, Version 4.5. Hannover, Germany: DataCite e.V., 2024. doi: 10.14454/g8e5-6293

[4] ISO/IEC 11179-1:2023. Information technology — Metadata registries (MDR) — Part 1: Framework

[5] R. L. Rodd, DOE Data Days 2023 Report, Apr. 2024. doi: 10.2172/2345124

Next Steps and Parting Thoughts

Next Steps

- xCards where x is {Data, Model, Code, Agent, Skill,...}
- Automated xCard generation as part of the data lifecycle
- Integrate xCards (or xCard content) into the data ecosystem.
- Leveraging the graph of digital objects
 - Provenance
 - Validation

Thoughts

- Web vs Hierarchy
- Smart data vs Intelligent Agents
- Catalogs on demand

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*

Thank you

data.jlab.org

JLabDataManagement@jlab.org