



Data Perspectives

AI-Ready and FAIR Data Across
the JLab EIC Community

October 02, 2026

Laura Biven

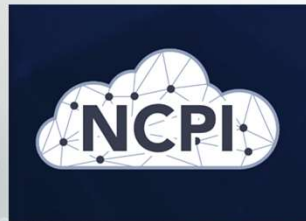
Chief Data Officer, Jefferson Lab

 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



My interests:

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

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:

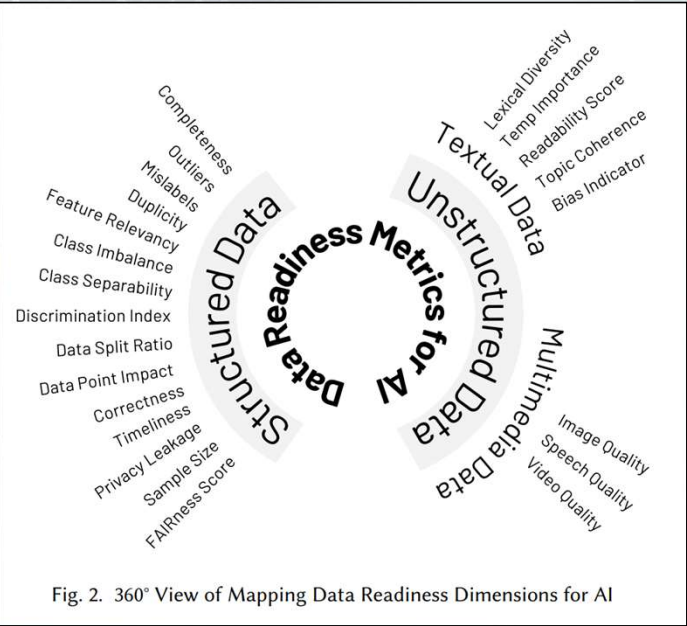
- Balance or representativeness (bias)
- Findability
- Ability to generate more data...

Data for Model Training

Data for Inference

MA Data for Agents

What Makes Data AI-Ready



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

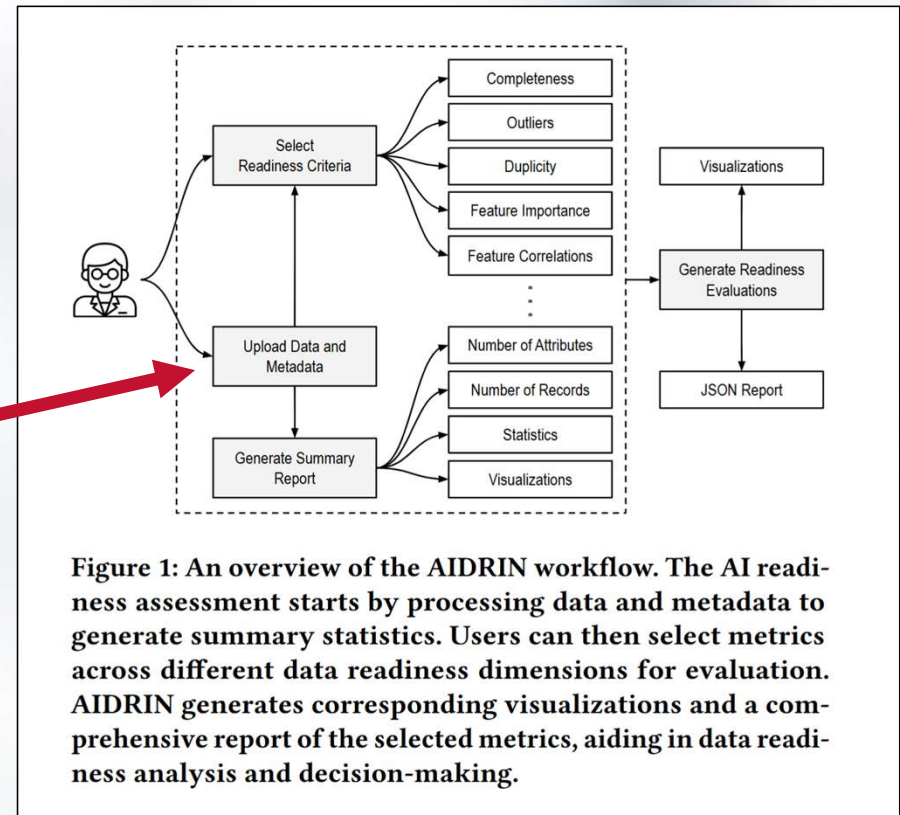
Table 2. 2D conceptual maturity matrix integrating Data Readiness Levels and Data Processing Stages (grey cells N/A)

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

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



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>

More about AIDRIN: <https://aidrin.org/>

Data Cards

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

Document Version: 1.0
Date: July 2020



Prepared by:
Diana McSpadden, Thomas Jefferson National Accelerator Facility, Newport News, VA, USA
ORCID: <https://orcid.org/0000-0002-8520-1831> | ROR: <https://ror.org/02vxxzd76>
Olga Kuchuk, Oak Ridge National Laboratory, Oak Ridge, TN, USA
ORCID: <https://orcid.org/0009-0000-2472-4710> | ROR: <https://ror.org/01x2mbd66>
Charlotte Neher, Sandia National Laboratories, Albuquerque, NM, USA
ORCID: <https://orcid.org/0009-0009-2169-0032> | ROR: <https://ror.org/01apw12>
Victor Walker, Idaho National Laboratory, Idaho Falls, ID, USA
ORCID: <https://orcid.org/0000-0001-5172-4410> | ROR: <https://ror.org/00y2u548>
Jean Luca Bez, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
ORCID: <https://orcid.org/0000-0002-3915-1135> | ROR: <https://ror.org/02b2y002>
Laura Biven, Thomas Jefferson National Accelerator Facility, Newport News, VA, USA
ORCID: <https://orcid.org/0000-0002-5755-8449> | ROR: <https://ror.org/02vxxzd76>

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

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: FAIR metadata + LLM-ready narrative + structured relationships

A compact, structured summary to collect, organize, validate, and present information needed to make digital resources **machine actionable** and **AI- and Agent-Ready**.

THE DATA PRODUCT

DATASET

files • tables • images • model outputs

The challenge: AI tools (and humans) **infer reuse guidelines** from incomplete metadata and publications **not originally intended** for that use.

Context, including tacit and community-specific knowledge, is key to assessing and improving AI readiness.

Agents must be able to **traverse the graph of digital objects**, e.g., datasets, software, AI models, publications.



THE DATA CARD

The data card combines **structured metadata (YAML)** and **computable natural language (Markdown)** to answer core questions:



What is it?

Purpose, scope, contents, variables



Where did it come from?

Source, instrument/model, provenance, version



How was it made?

Methods, processing, calibration, code/workflow



Can I trust it?

Quality checks, uncertainty, known limitations



Can I use it?

Access, rights, sensitivities, restrictions



How to I navigate the web of scientific information?

Identifiers & structured relationships, schema, formats.

In this hackathon, a data card serves as the perception layer.

It lets the next team, model, or agent quickly answer:

Relevant?

Is this the right data? What /where are the related resources?

Usable?

Can I legally + technically use it?

Meaning?

How do I interpret the data?

Reliable?

What should I trust or question?

Reusable?

Can I connect it to my workflow?

Not a replacement for full metadata or documentation — a data card is the concise entry point that makes those resources discoverable and interpretable.

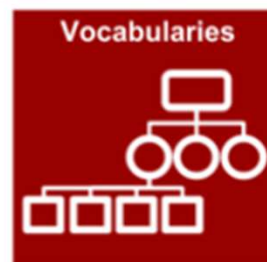
Schema, template, field guide & tools: doi.org/10.18141/3376151
Technical report: doi.org/10.2172/3377514
Data card generation skill: <https://github.com/AI-ModCon/genesis-skills>

Persistent, Unique Identifiers

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>

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

Data Governance – Review (in development)

Data Card

Jefferson Lab Data Review and Govern Checklist

The purpose of this document is to ensure all appropriate policies and requirements are considered before data and other digital artifacts are released outside of the lab, which includes submitting data to outside repositories and sharing with external collaborators. The result of this review is a determination regarding whether and how data should be released outside the lab, including any rules for access (e.g. controlled access) and reuse (e.g. licensing type). Distinct determinations may apply to the metadata, data card, and the underlying data asset.

Governance refers to both the conditions for access and the conditions for reuse. Governance rules often flow from a data set to any resulting derived products including derived data or AI models. Data derived from multiple datasets will typically inherit the most restrictive access and reuse rules of its parent artifacts. Governance rules must also follow the award or contract terms that apply to the research activity for which the data were generated or collected.

It's important for this process to be complete and as streamlined as possible to meet increasing expectations for data sharing. For this reason, this review process may be carried out for data before they are generated as long as an accurate and detailed data card can be generated.

The checklist serves as both a tool for reviewing a dataset against the listed policies and requirements as well as the archived documentation of the determination. The checklist should contain all of the information, or references to information, supporting the final

Defines a policy-based
due diligence review

Updated Data Card with
Governance Rules

This review framework is
coordinated across labs
through GM

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 for
 - Generative workflows
 - Provenance
 - Validation

Thoughts

- Web and Hierarchy
- Smart data and Intelligent Agents
- Guardrails

Summary

- There's no recipe for AI-Readiness
- Facets of AI-Readiness include
 - Structured, FAIR metadata
 - Contextual information
 - Clear, machine actionable governance
- Some areas of ongoing research
 - Creating accurate metadata and context at scale as part of the data lifecycle
 - Integrating data card information into the data ecosystem
 - Machine-actionable data and cybersecurity
 - ...

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