

LDRD LD2615 · JLab EIC Friday meeting · 2 October 2026

FAIR and AI-Ready Data for CEABAF

A case study of the JLAB data ecosystem

Anil Panta, Brad Sawatzky, Casey Morean, Dmitry Romanov, Douglas Higinbotham
Thomas Jefferson National Accelerator Facility

Why FAIR and AI readiness, and why now

DOE policy

DOE Order 241.1D and public-access requirements expect outputs to stay discoverable and reusable.

DOE AI initiatives

The Genesis Mission and American Science Cloud link data, computing, facilities, and AI.

JLab goal

The Office of the Chief Data Officer has set a FAIR, AI-ready data ecosystem as the lab's goal.

"Data" here means every artifact: raw and reconstructed events, simulation, calibrations, run conditions, software, workflows, and documentation.

Our working definition: two layers

LAYER 1

FAIR

Can people and machines find the data, get it, understand it, and reuse it? This is about the data's packaging.



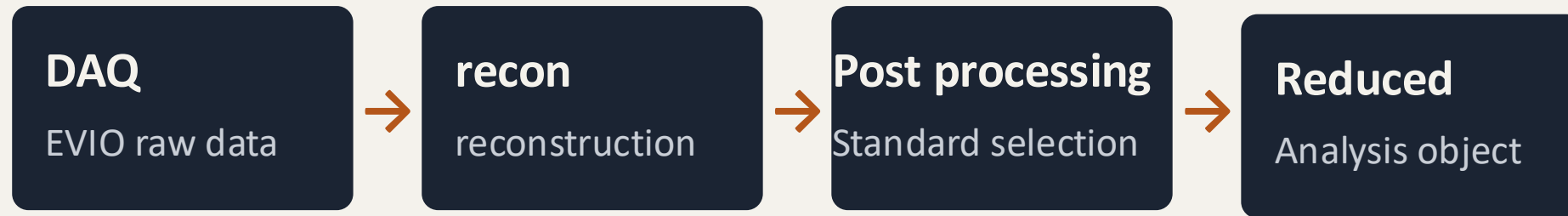
LAYER 2

AI readiness

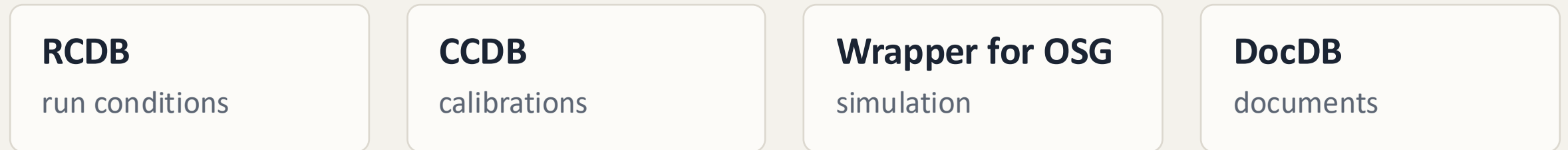
Is the content good enough for AI to rely on? Complete, well documented, with known uncertainties and no hidden biases.

We argue that data must be FAIR first to be AI ready. And AI readiness has separate criteria on top of FAIR.

Pilot: the standard data chain in CEBAF



SUPPORTING SERVICES



What FAIR means

F

Findable

Can it be found?

- F1** Permanent, unique ID
- F2** Rich description
- F3** Description names the ID
- F4** Listed in a searchable catalog

A

Accessible

Can it be retrieved?

- A1** Fetched by ID with a standard protocol
 - A1.1** Protocol is open and free
 - A1.2** Login and permissions where needed
- A2** Description outlives the data

I

Interoperable

Can it be combined?

- I1** Shared, formal language
- I2** Shared vocabularies
- I3** Links to related data

R

Reusable

Can it be used again?

- R1** Accurate, detailed description
 - R1.1** Clear usage license
 - R1.2** Full history of how it was made
 - R1.3** Meets community standards

Wilkinson et al., Scientific Data 3, 160018 (2016). FAIR does not require open data, only clear access rules.

FAIR assessment:

0

met

5

partial

10

not met

of the 15 FAIR sub-principles,
at the time of assessment

No persistent identifiers or catalog for data products. Discovery relies on file paths and expert knowledge.

RCDB and CCDB describe runs and calibrations, not the files that carry the data.

Processing context lives in paths, shell arguments, and database rows, not as provenance.

FAIR findings: where GlueX stands

PRINCIPLE	WHAT WE FOUND	STATUS
Findable Data and metadata have permanent IDs and are indexed in a searchable catalog	- No persistent IDs for raw, reconstructed, or analysis data. - Metadata exist at run level in RCDB and CCDB but are not attached to files. - Discovery relies on directory layout and expert knowledge.	F1, F3, F4 not met F2 partial
Accessible Data and metadata can be fetched by ID with a standard, secure protocol	- Access is by file path on JLab systems, controlled by UNIX groups. - No dataset-level retrieval protocol. - RCDB and CCDB metadata persist but no obvious link to data	A1, A1.1 not met A1.2, A2 partial
Interoperable Data and metadata use shared formats, vocabularies, and links	- Formats are well structured, but meaning lives in code. - Processing context sits in scripts, arguments, and database rows, not as provenance.	I2, I3 not met I1 partial
Reusable Data and metadata carry descriptions, licenses, provenance, and standards	- No per-dataset metadata or usage terms. - Provenance is incomplete - containers preserve the environment but not the workflow	R1, R1.1, R1.3 not met R1.2 partial

FAIR solutions: what we propose

PRINCIPLE	SOLUTION FROM THE PAPER
Findable Finding 1	- Rucio as data and metadata catalog: permanent IDs for data each with queryable metadata. - Run collections make analyses citable and can be posted in InvenioRDM
Accessible Findings 1 and 2	- Fetch data by Rucio ID through standard protocols (XrootD, WebDAV) with access control. - Rucio ID accessible via metadata - RCDB run records link to catalog IDs.
Interoperable Findings 2 and 3	- Common metadata fields across halls; - machine-readable interfaces for RCDB and CCDB. - Provenance captured when jobs run, using the W3C PROV model stored as JSON-LD.
Reusable Finding 4	- InvenioRDM for documents, software and data records, with DataCite metadata and DOIs through OSTI. - Software environment and workflow capture

Where the FAIR fixes stand today

RUNNING

Data catalog

Rucio gives data permanent IDs.
Serves ePIC; GlueX in pre-prod; clas12 in progress

RUNNING

Document/Data repository

InvenioRDM deployed; all DocDB documents migrated.
InvenioRDM as Data repository (Genesis) underway

PROTOTYPE

Automatic registration

Files get an ID and history the moment the DAQ writes them.

PROTOTYPE

DOIs for documents

Repository issues DOIs through OSTI; tested.

IN DEVELOPMENT

Linked systems

Catalog, SWIF jobs, and repository share one ID system, with CST.

UNDER EVALUATION

PanDA

Large-scale job management with workflow capture

Open data: one path from catalog to public release



AVAILABLE IN RUCIO

Rucio supports open data, so released datasets need no copy or separate system.

UNDER DESIGN

Rucio and InvenioRDM integration, so an open dataset gets a citable public record.

PRINCIPLE

FAIR does not require open data, but release should be one step when allowed. Adheres to Open data OSTP policy

AI readiness:

Is it complete and correct?

No missing runs, duplicated events, or malformed values.

Do we know where it came from?

Which calibration, software version, and settings produced it.

Do we know how precise it is?

Detector resolutions and systematic uncertainties for each run period.

Does it represent what it should?

Rare channels and trigger acceptance are properly accounted for.

Our framework turns these questions into 26 measurable checks in seven categories. Details are in the preprint.

The 26 AI readiness checks on GlueX data product

1. Quality

- Completeness
- Anomalies
- Duplication
- Data preparation
- Timeliness

2. Understandability

- Metadata quality
- Provenance
- Data access interfaces

3. Structure

- Data types
- Data schema
- Storage and formats
- Access performance

4. Value

- Significance
- Ground truth
- Sample influence
- Uncertainty

5. Fairness and bias

- Class imbalance
- Class separability
- Group disparity
- Representation

6. Governance

- Collection
- Curation
- Application
- Security
- Privacy

7. Application

- Model metrics

Can existing tools do these checks? Only partly.

10/26

checks done well by at least
one of the three tools

Great Expectations

Checks data against rules you write, such as allowed ranges and required fields.



deepchecks

Tests datasets and machine-learning models for outliers and shifts between samples.



AIDRIN

Scores scientific datasets for AI readiness, but mostly checks that metadata fields are filled in.



What existing tools miss

Not covered well by any tool

- Where the data came from
- Uncertainty on each measurement
- Whether a few events drive a result
- How well signal separates from background
- How current and how fast to access the data are

Built for spreadsheets

Event data must be flattened into tables first, which loses their structure.

Tuned for industry

One tool rejects data where a class is under 10% of another, which is normal for rare physics processes.

Next frontier: getting ready for AI assistants

Look up

Assistants can search catalogs and documents freely, within each user's permissions.

Act

Changes go through defined interfaces with temporary permissions, and people approve anything important.

Record

Every important action records what information was used and who approved it.

Principle: read broadly, act narrowly, record everything.

Summary and next steps

What we found

- FAIR is needed, but it is not enough for AI.
- Metadata are there but unlink to data and no single interface
- Uncertainties are not recorded as data.
- Existing tools do 10 of 26 checks well.

What comes next

- Data policy recommendation to Data Office.
- Add manpower to help with integration to exp.
- Training the experiment data manager
- Explore more of solutions and start implementation.

Preprint: [arXiv:2609.34987](https://arxiv.org/abs/2609.34987) · Contact: panta@jlab.org

Thank you

Preprint: [arXiv:2609.34987](https://arxiv.org/abs/2609.34987)

Contact: panta@jlab.org

LDRD LD2615 · Thomas Jefferson National Accelerator Facility

Backup: six findings and their status

Finding	Remedy	Status
1. Datasets have no permanent IDs	Central data catalog (Rucio)	Running for ePIC; GlueX in progress
2. Metadata serve software, not users	Register files at creation; common fields	Registration prototyped
3. Processing history not recorded	Record it automatically as jobs run	Prototype
4. Documents and code hard to find or cite	InvenioRDM repository with DOIs	Repository running
5. Uncertainties not recorded as data	Curated uncertainty tables in CCDB	Proposed
6. Tools miss physics-critical checks	Build physics-specific checks	Proposed