

Genesis Mission through the Data Lens

Digital Twins for Nuclear and Particle Physics (NPTwins 2026)

5/11/2026

Laura Biven

Chief Data Officer, JLAB

Jefferson Lab

U.S. Department of ENERGY

Laura Biven

Chief Data Officer



Laura Biven joined Jefferson Lab in 2024 as the Chief Data Officer, leading data strategy in support of the Lab and DOE missions. She co-leads data efforts for the Genesis Mission.

Prior to joining JLab, Laura spent over 15 years in US government funding agencies (DOE and NIH) leading programs in data management, AI, and data science infrastructure.

Laura holds a first class MSci in math and physics from the University of Bristol and a Ph.D. in applied mathematics from the University of Warwick, both in the UK.





PuRe Data Resources

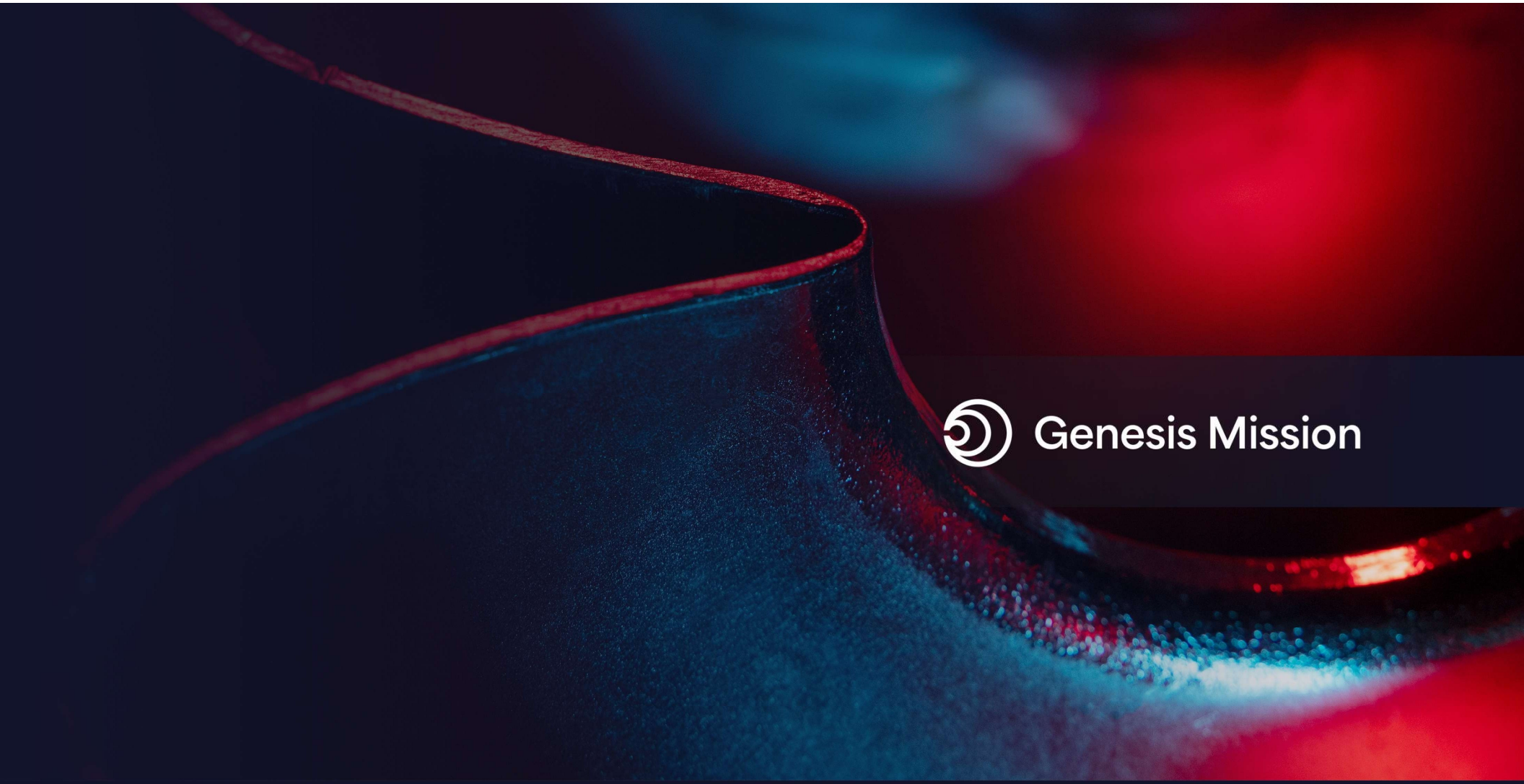
NCPI





AIM AHEAD

data.jlab.org
orcid.org/0000-0002-5755-8449



U.S. DEPARTMENT of ENERGY

LLNL-PRES-2016998

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



Jefferson Lab

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.



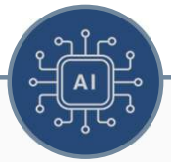
Infrastructure

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.

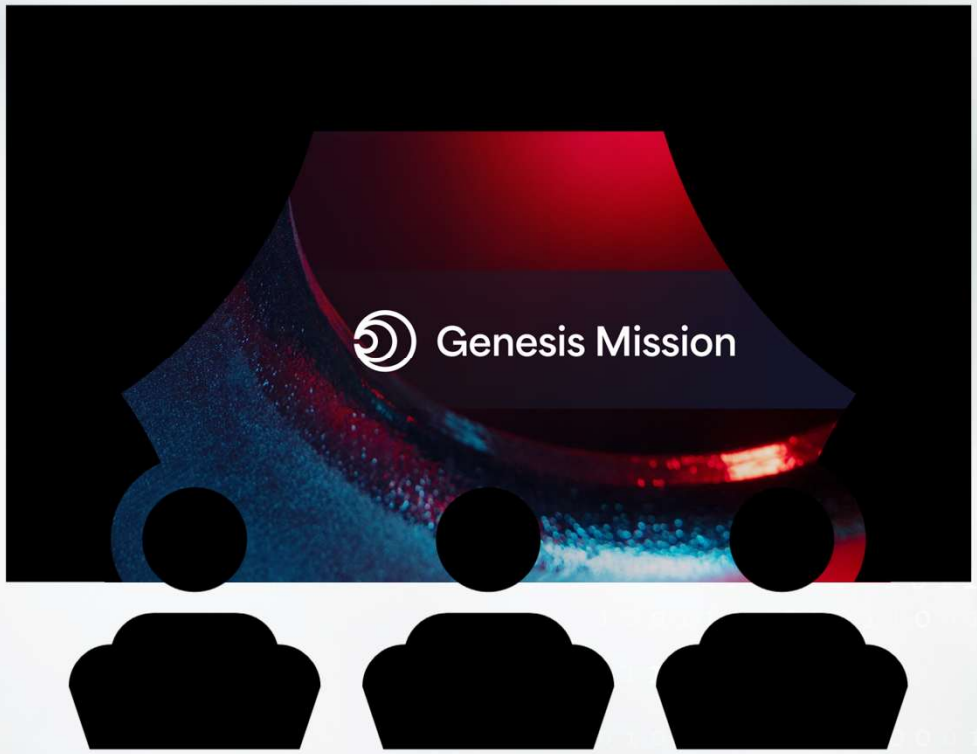
Genesis Mission

Bringing Together World-Class Innovators Across Sectors

Under the leadership of DOE and the White House Office of Science and Technology Policy, the Genesis Mission brings together DOE's National Laboratories with America's leading universities and industry, including pioneers in artificial intelligence, computing, materials, and energy.



End of Official Genesis Slides



Genesis Mission Goals- through the data lens

GOAL: assemble the world's largest and highest-quality scientific data sets

- What is AI-Ready?
- What is Machine Actionable for autonomous, agentic workflows?
- How will this collection be stewarded?

GOAL: dramatically accelerate scientific discovery

Overcome current rate-limiters to scientific productivity and security:

- Findability and accessibility of data and related artifacts for people and agents
- Lack of interoperable, secure, trustworthy, frameworks for data sharing and publication
- Manual data curation and AI-readiness

Data Management for the Future of Science



Image generated with CoPilot

Agents working on behalf of researchers, including personalized digital twin research assistants. Agents will need to discover, assess the fitness of data, access, and reuse data with minimal or no intervention from humans.

Sharing and reuse of {data, computing, infrastructure…} by a wide variety of stakeholders. Detailed information about resource governance and access requests will need to be compared, with trust relationships at scale. (Governance covers rules for access and reuse – who, what, when, where, how, purpose.) (Sovereignty of resources, subject to policy).

Optimized workflows (by agents or otherwise) that take into account time to solution, cost, safety, and scientific impact. Such decisions will be supported by operational data about a resource's availability, cost, power consumption, performance,… and experiential data generated by previous workflows.

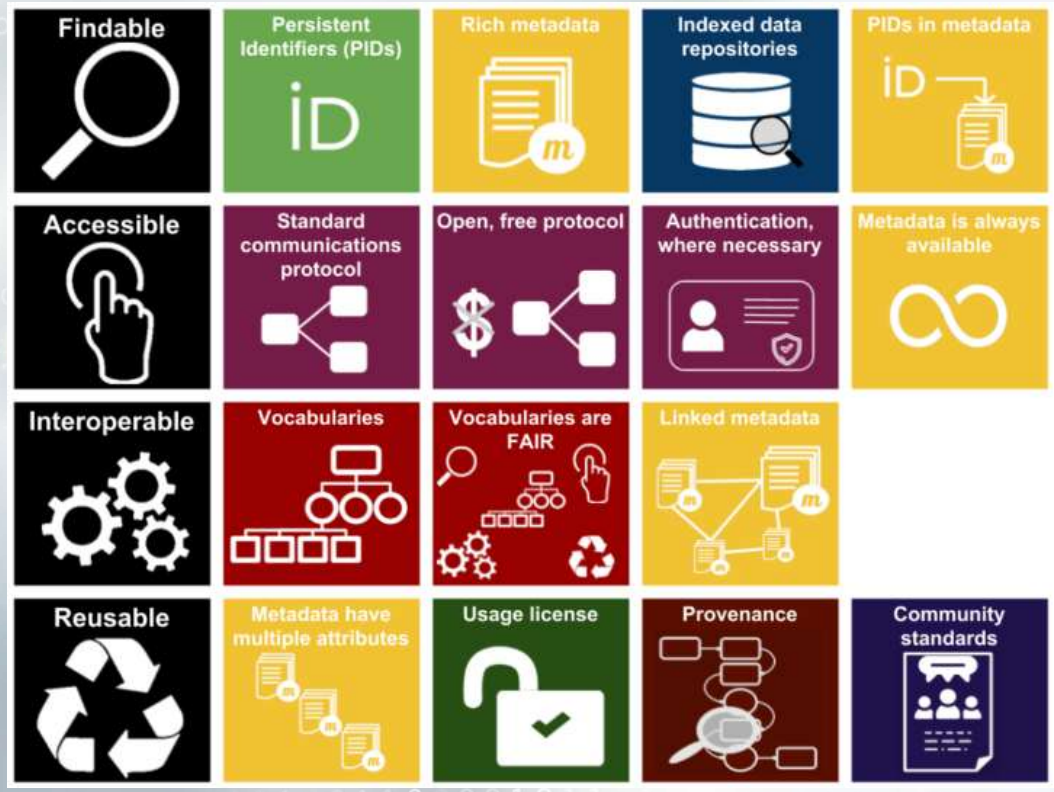
The ability for humans to be creative and innovative, and the right to personal privacy will remain paramount. Data management and infrastructure securely manages PII and digital twin assistants and preserves researcher ability to be creative and innovative.

Reusable artifacts must be stewarded long-term with impact metrics over time. i.e. Long-term stewardship is a necessary condition for reusability.

Protection for the integrity of high-value digital assets is a cornerstone of scientific productivity / gold standard science. E.g., protect against pollution, corruption, bit rot,…

FAIR Data → Smart Data

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)

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

FAIRification

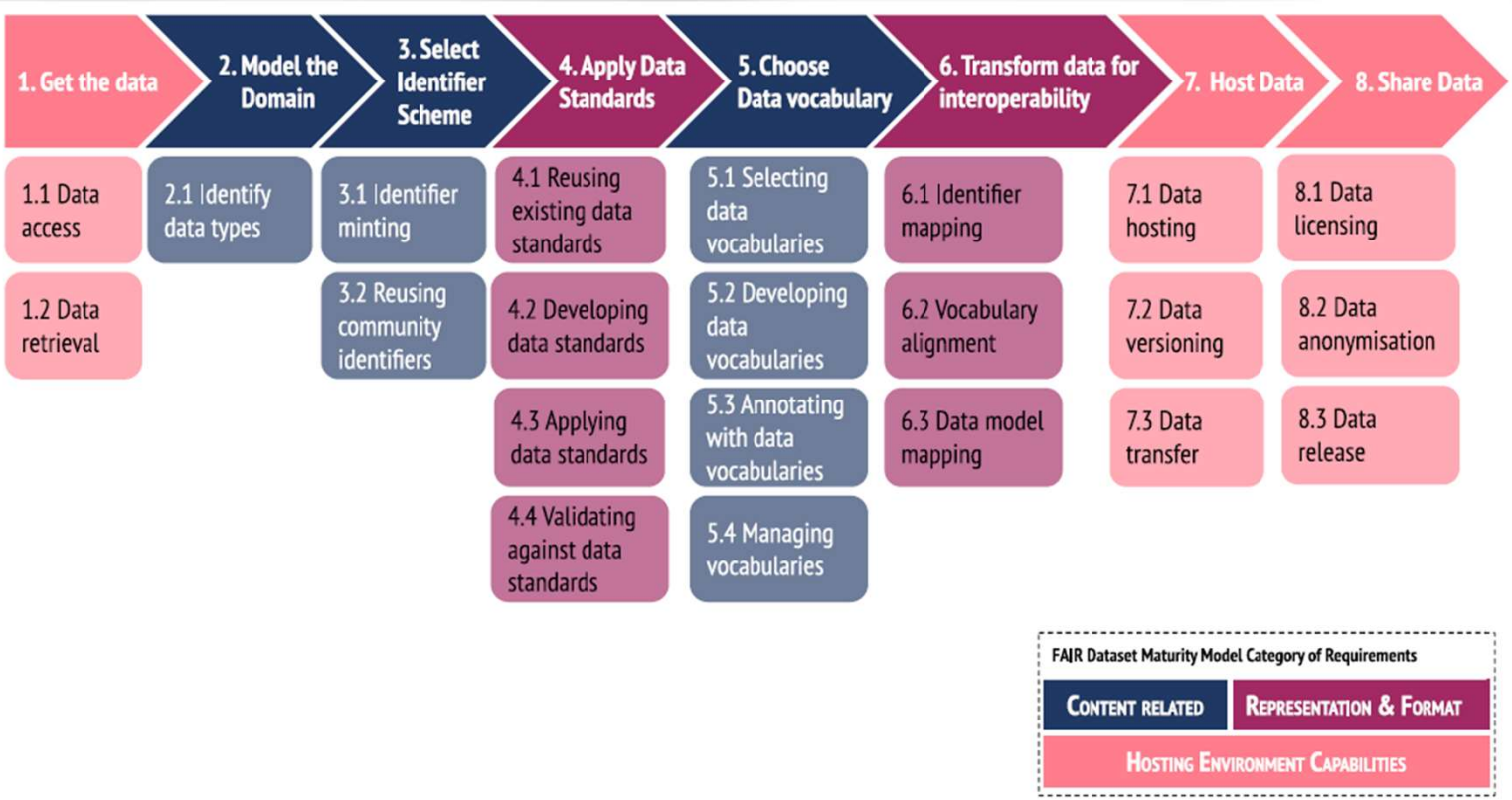


Fig. 4 The FAIRification template steps. Each step is colour-coded based on whether its implementation applies to data hosting, representation and format or data content. Each step is broken down into one or more sub-steps. More details can be found in Supplementary Table 2.

Welter, D., Juty, N., Rocca-Serra, P. *et al.* FAIR in action - a flexible framework to guide FAIRification. *Sci Data* **10**, 291 (2023). https://doi.org/10.1038/s41597-023-02167-2

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

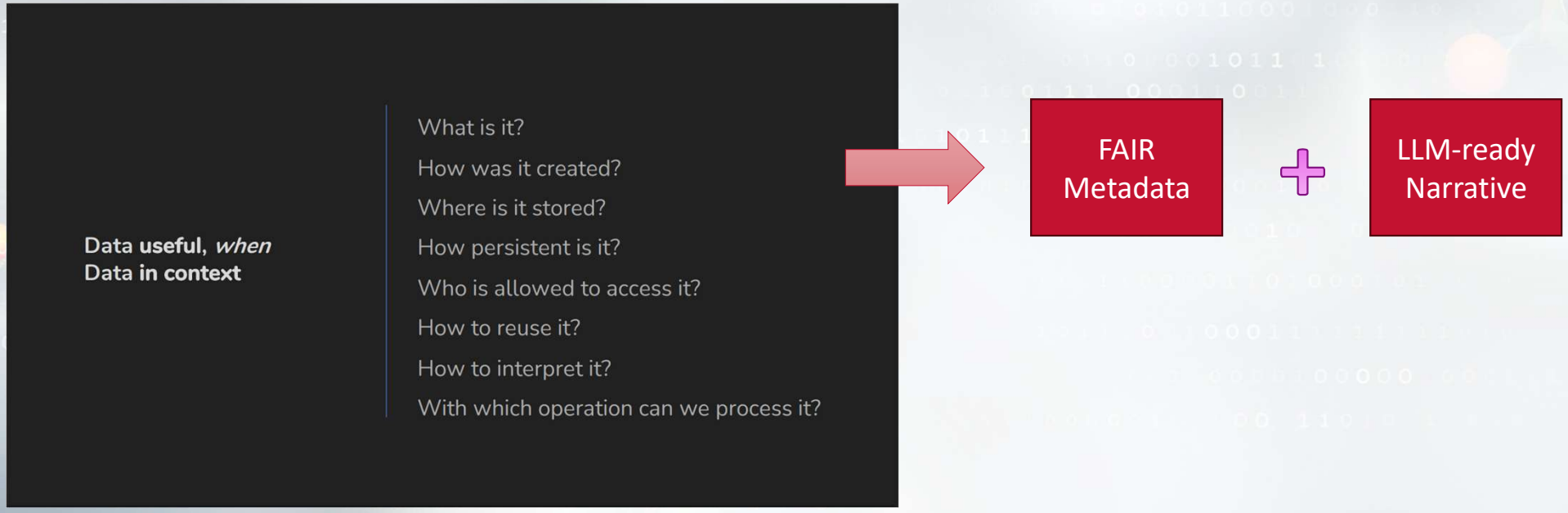
FAIR training Materials https://doi.org/10.1371/journal.pcbi.1007854

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 Context for machine actionability



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

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

FAIR DIGITAL OBJECTS

https://fairdo.org/

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

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

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. arXiv. https://arxiv.org/abs/1803.09010

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

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 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:
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.

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:

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

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:

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:

PIDs + xCards + provenance enable reproducible workflows

Workflows themselves are training data for GenAI capabilities

AI-Ready Data

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…

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***

Jefferson Lab

Thank you

data.jlab.org

JLabDataManagement@jlab.org

U.S. DEPARTMENT of ENERGY