

FAIR and AI-Ready Data in ePIC

The Data Foundation for the ePIC Computing Model and Its Closed-Loop Scientific Workflows

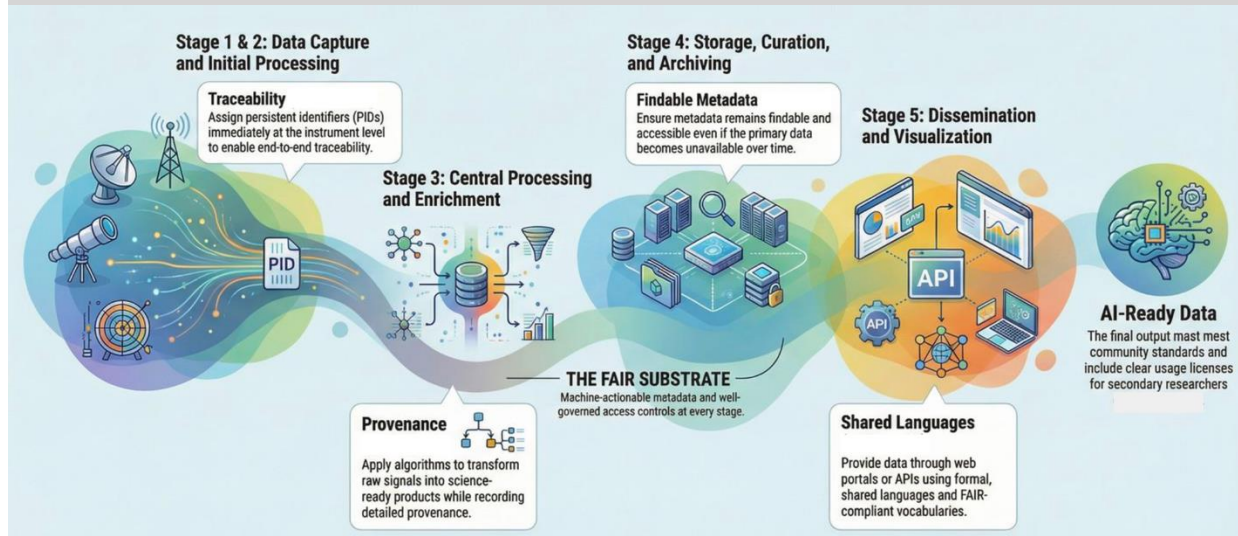
Markus Diefenthaler (Jefferson Lab)

AI Readiness for the ePIC Detector and FAIR

Definition: AI Readiness in the context of the ePIC experiment refers to the extent to which the detector, its computing, data, and the associated workflows can support the development, integration, and sustained application of AI technologies.

AI readiness is not just whether we can build an AI model. It is whether ePIC can integrate and operate AI reliably as part of the experiment over time.

Connection: FAIR provides the data foundation for AI readiness. AI readiness builds on FAIR by adding the capabilities needed to for sustained, reliable AI operation.



FAIR is not only about the data. It captures the **data, metadata, and provenance** throughout the full data lifecycle.

In ePIC, **FAIR and AI readiness support scientific workflows**, including **closed-loop workflows**, where information from processing and analysis **feeds back** to improve subsequent detector operation, data taking, and data processing.

The ePIC Computing Model Enables Closed-Loop Scientific Workflows

The opportunity is not just to use AI, but to rethink the scientific workflow.

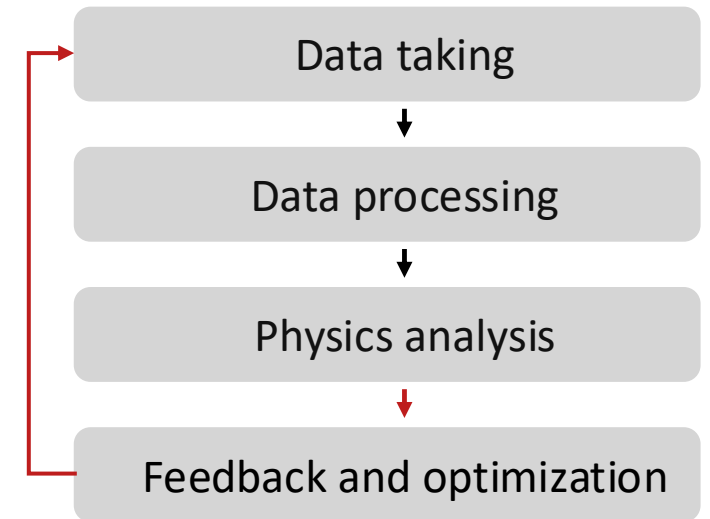
"The AI revolution in science is a workflow revolution."

David Lawrence, EICUG-ePIC meeting

Streaming readout, AI, and distributed computing provide the foundation for this transition.

- 1 Streaming Readout**
Maximizes scientific reach through unbiased event selection and continuous measurement of backgrounds and detector noise.
- 2 AI**
Accelerates science through autonomous alignment, calibration, and validation, shortening the path to physics results and publications.
- 3 Distributed Computing**
Enables flexible access to dedicated and opportunistic computing resources where they are needed.

Toward a Closed-Loop Workflow



AI Is Already Part of the ePIC Workflow

epicprod

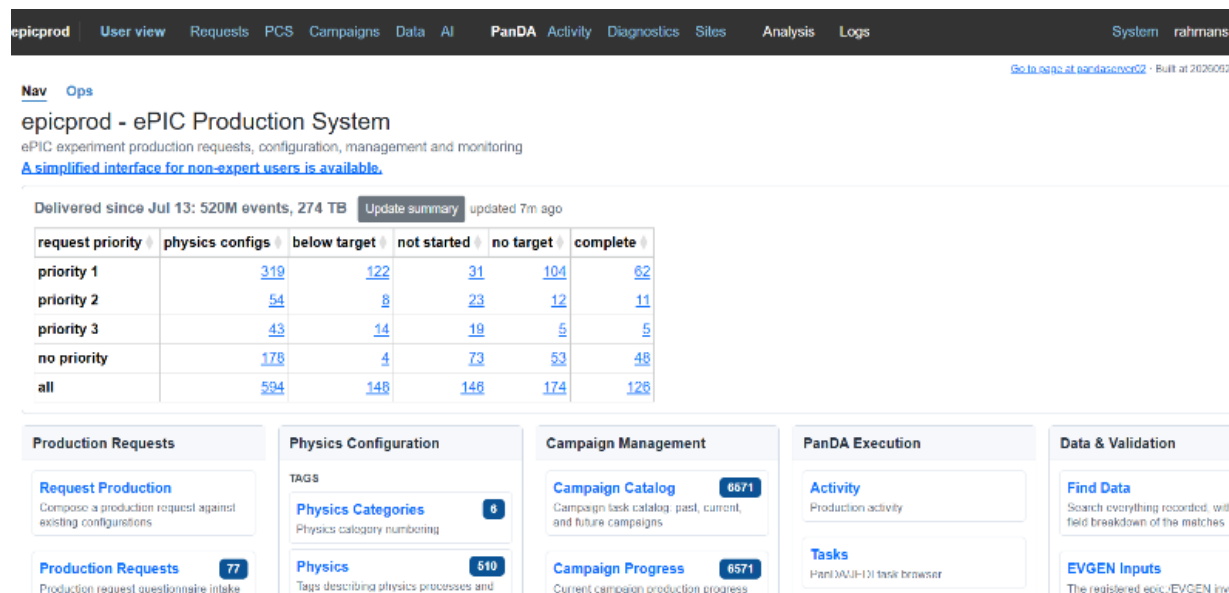
Autonomous system for simulation production in ePIC, based on PanDA and agentic workflows.

<https://epic-devcloud.org/prod/>

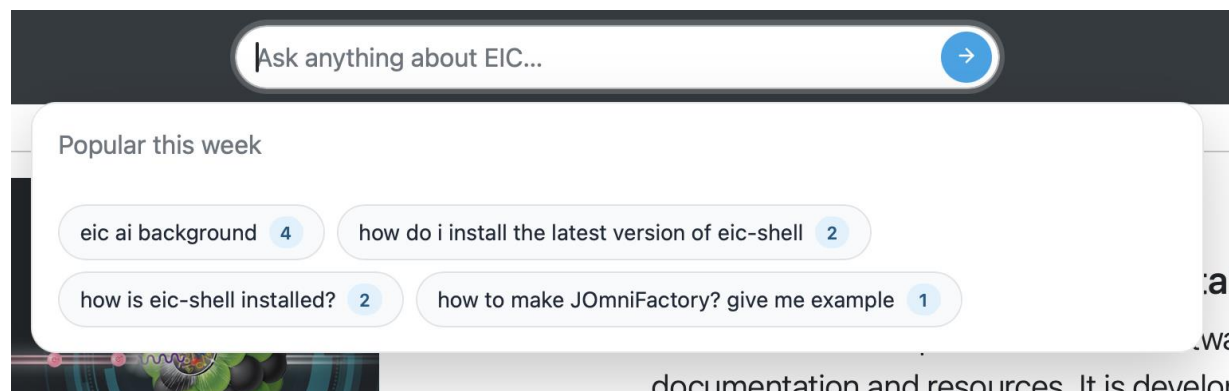
AI-enabled search

Supports onboarding and user learning in ePIC through AI-assisted search across documentation and tutorials.

<https://eic.github.io>



request priority	physics configs	below target	not started	no target	complete
priority 1	319	122	31	104	82
priority 2	54	8	23	12	11
priority 3	43	14	19	5	5
no priority	178	4	73	53	48
all	594	148	146	174	126



FAIR as the Data Foundation for ePIC

Findable

Discoverable data products through **structured, queryable metadata in Rucio** and persistent identifiers. Findability extends across the data lifecycle so that datasets and their associated context can be located by both people and automated workflows.

Accessible

Data access through **Rucio for managing data locations and replicas, and XRootD for data access and transfer**. ePIC provides open access for reading its data, **with data also available as flat files for straightforward use**.

Interoperable

Interoperability is built around a common event data model, **EDM4eic**, with its **structure and relations defined in a machine-readable YAML specification**. Interoperability also includes integration with the **broader HENP ecosystem**.

Reproducible, Reusable, and Reinterpretable

Software, containers, configuration, and other context needed to interpret and reproduce scientific results are preserved. **Data-card-driven reproducible workflows** are being introduced to capture the information needed to recreate datasets and results.

FAIR turns ePIC data and its scientific context into a foundation that people, workflows, and AI can reliably use.

FAIR Ingredient: Executable Provenance

```
10 # for campaign datasets, rucio DID provides
11 # a readymade Identifier
12 identifier:
13     "epic:/REC0/26.07.1/epic_craterlake/DIS/pythia8.316-1.0/MC/noRad/ep/10x100/q2_100to1000"
14
15 # short human-readable description of
16 # the dataset
17 description:
18     "26.07.1 MC DIS sample with Q^2 = 100 - 1000 GeV^2 generated by Pythia8.316"
```

(1)

```
11 configfile: "config.yml"
12
13 # prestart: generate input, rules -----
14
15 import cardio
16 tcard = cardio.load_card(config['template'])
17 icard = tcard.input()
18 tcard.dump_rules(config['rules_file'], config['config_file'])
19 icard.dump_files(config['input_list'], protocol="rucio")
20
21 # startup: make output dirs -----
22
23 onstart:
24     import os
25     os.makedirs(config['hist_out'], exist_ok=True)
26     os.makedirs(config['plot_out'], exist_ok=True)
27
28 # run workflow -----
29
30 include: config['rules_file']
31
32 rule all:
33     input:
34         f"{config['plot_out']}/matchedJetResolutionVsEta.{config['suffix']].png"
35
36 # shutdown: write output card -----
37
38 onsuccess:
39     import cardio
40     card_name = config['hist_out'] + "/" + config['card_name']
41     cardio.make_card(card_name, config['template'], config['hist_out'])
```

(3)

```
43 configuration: |
44     run_dir:
45         "run"
46     hist_out:
47         "out"
48     plot_out:
49         "out/plot"
50     suffix:
51         "demo"
52     workflow: |
53         rule make_hists:
54             input:
55                 list = f"{config['run_dir']}/input.list",
56             output:
57                 hist = f"{config['hist_out']}/output.root",
58             params:
59                 nfile = 3,
60                 nevent = 100,
61             shell:
62                 "root -b -q analysis/MakeJetValidationHists.C '\{output.hist}\', '\{input.list\}'"
63         rule make_plots:
64             input:
65                 hist = f"{config['hist_out']}/output.root",
66             output:
67                 plot = f"{config['plot_out']}/matchedJetResolutionVsEta.{config['suffix']].png"
68             params:
69                 path = f"{config['plot_out']}",
70                 suff = f"{config['suffix']}",
71             shell:
72                 "root -b -q analysis/MakeJetValidationPlots.C '\{params.path}\', '\{params.suff\}'"
```

(2)

1) Key entries from example input card
2) Workflow description from output card
3) Snakefile using cardio to run workflow from output card

Datacard: Information needed to recreate a dataset or plot.

Workflow: Snakemake + cardio extension creates or consumes the datacard.

Reproduce: Rerun from the output datacard to reproduce the result.

Provenance becomes reusable when it can drive the workflow that created the result.

AI Readiness Beyond FAIR

1. The availability of **high-fidelity simulated datasets** for training and validation of AI models.
2. **Experimental data** (intended in a broad sense to include data from the detector-computing system and associated data streams such as slow control data and logbooks) for AI inference and validation.
3. Availability of **inference engines**, deployable at the edge (ASICs and DAM/FPGA boards) and within the computing ecosystem (CPUs, GPUs) to achieve latency requirements.
4. **Uncertainty quantification and interpretability**: Guardrails and fallback policies when models are uncertain and/or inputs are OOD.
5. **MLOps**: Sustainability of AI models relative to data drift, out of distribution data, etc. Clearly defined ownership of AI services, including responsibility for model development, deployment, monitoring, documentation, and incident response.
6. **Control interfaces and actuation**: Interfaces that enable validated AI outputs to safely modify detector, DAQ, or computing-system parameters as part of a closed-loop workflow.
7. **The human factor**: designing for humans in the loop in an attention-aware manner. By attention-aware, we mean in the sense that humans cannot physically monitor all things continuously. Systems can and should be designed around this fact to augment human capabilities.

FAIR provides the data foundation. Beyond FAIR, AI readiness adds requirements for sustained operation, including inference, uncertainty, control interfaces, MLOps, ownership, and human oversight.

FAIR and AI Readiness Enable Closed-Loop ePIC Science

FAIR Data Foundation

Findable • Accessible • Interoperable • Reusable

Data and metadata, the scientific context, that humans, workflows, and machines can reliably use.

AI Readiness

Data • Inference • Uncertainty • MLOps • Interoperability • Human Oversight

The capabilities needed for sustained, reliable AI across the experiment.

Closed-Loop Scientific Workflows

Data Taking → Data Processing → Physics Analysis → Feedback and Optimization

An integrated workflow connecting data, computing, AI, and humans to accelerate and maximize science.

Together, FAIR and AI readiness provide the data foundation for closed-loop ePIC science.