Datasheets for RHIC

11/20/25

What Are Datasheets?

- Structured, human-readable documents that describe a dataset.
  - The concept of "Datasheets for Datasets" dates from 2018,
  - In AI/ML world it's been used to document datasets for transparency, reproducibility, and fairness.
- Datasheets capture essential information like (adapted to our case):
  - Purpose and physics motivation
  - Data production workflow
  - Detector configuration & conditions
  - Known limitations and caveats
  - Data quality, provenance, and responsible contacts
- They serve as the narrative companion to a dataset

How Datasheets Relate to Metadata ?

- Metadata = machine-readable descriptors
  - files, runs, luminosity, detector conditions, software versions, etc.
- Datasheet = human-readable context
  - Interprets and organizes the metadata
- Datasheets include metadata and also add:
  - Explanation of the data's purpose and scope
  - Interpretation of metadata values
  - Constraints and known issues not captured in metadata
  - Guidance for analysts on the correct use of the dataset

Why Datasheets for RHIC DAP ?

- Support long-term usability of RHIC datasets beyond the lifetime of active collaborators
- Provide a reference for:
  - What a dataset contains,
  - How it was produced, how it has been used,
  - What is safe, correct, or recommended for analysis
- Facilitate reproducibility by documenting assumptions and caveats
- Support FAIR principles
- Needed for future re-analysis, AI-supported analysis, and cross-experiment comparison
- Help prepare RHIC data for AI-ready in support of the AmSc initiative

Examples of What a RHIC Dataset Datasheet Could Include

- Dataset identifier, run period, data tiers (RAW, QA, DST, skim)
- Trigger logic and detector configuration
- Calibration and reconstruction versions
- Data quality summary and bad run lists
- Recommended analysis software/workflows
- Known systematic issues
- POCs responsible for maintenance or historical knowledge

Datasheets and the SciBot

- SciBot is an AI assistant that's
  - Answers questions about datasets, run conditions, and analysis workflows
  - Guides users in exploring and interpreting complex RHIC data

- Datasheets serve as a knowledge base for SciBot, ensuring answers are accurate, reproducible, and traceable

- SciBot uses datasheets to provide:
  - Physics motivation and dataset scope
  - Detector and trigger configuration
  - Known limitations or caveats
  - Recommended analysis procedures

Datasheets stored in InvenioRDM

- InvenioRDM:
  - Centralized storage for documentation in one platform
  - FAIR-aligned
  - Supports versioning & provenance (track updates)

- InvenioRDM is the natural place to store datasheets

Datasheet adaptation for RHIC context

AI/ML Datasheet Element	RHIC Adaptation
Purpose / Motivation	Physics analysis goal, detector used, run period
Collection Process	Trigger settings, beam energy, detector configuration
Preprocessing / Cleaning	Reconstruction algorithms, calibration, QA cuts
Composition & Statistics	Number of events, luminosity, dataset tiers (RAW, DST, skim)
Limitations / Biases	Bad runs, detector inefficiencies, acceptance effects
Ethical / Licensing	Collaboration rules, authorship, usage restrictions

RHIC Dataset Datasheet Template - v1

Dataset Name: [Insert dataset name]
Version: [Insert version, e.g., 1.0]
Maintainer / Contact: [Name, email, institution]
Repository / DOI: [Internal path or public link]
Funding / Acknowledgments: [List funding agencies, collaborations, or individuals to acknowledge]

1. Motivation / Purpose

- **Scientific Goal:** [Describe the physics question or goal addressed by the datasete.g., "Study of quark-gluon plasma properties in Pb-Pb collisions at 5.02 TeV."]
- **Intended Use:** [List primary use cases, e.g., "Event reconstruction, machine learning training, cross-section measurements, or detector performance studies."]
- **Not Intended For**: [Prohibited or unsupported uses]

2. Composition

- **Number of events / samples:** [Number of events or samples]
- **Data types:**
  - Event-level: [e.g., run ID, event ID, centrality, trigger type]
  - Track-level: [e.g., px, py, pz, momentum, charge, track quality]
  - Particle-level: [e.g., PID, mass, dE/dx, TOF info]
  - Vertex-level: [e.g., primary vertex x, y, z]
- **File formats:** [e.g., ROOT, HDF5, CSV]
- **Units:** [Specify units for all numeric fields]
- **Data volume**: [e.g., "500 GB (compressed)"]

3. Collection / Acquisition Process

- **Detector / Experiment:** [STAR TPC, TOF, ZDC, etc.]
- **Run period:** [Start date – end date]
- **DAQ configuration / Trigger setup:** [Details of triggers, luminosity, readout]
- **Simulation software (if applicable)**: [e.g., "PYTHIA XX, HIJING YY"]
- **Event selection criteria:** [e.g., minimum bias, vertex cuts]
- **Preprocessing / Reconstruction:** [Steps applied to raw data to produce DST]

4. Preprocessing / Cleaning / Annotation

- **Track-level cleaning:** [e.g., remove tracks with < N hits, poor fit]
- **Vertex-level cleaning:** [e.g., z-position within detector acceptance]
- **Particle-level annotations:** [e.g., PID, mother/daughter IDs, quality flags]
- **Excluded data:** [e.g., events with zero primary particles]

5. Known Limitations / Biases

- **Physics model limitations**: [e.g., "HIJING does not include hydrodynamic flow for low-energy collisions"]
- **Detector limitations**: [e.g., "Tracking efficiency drops below 80% for pT < 200 MeV/c"]
- **Quantitative biases**: [e.g., "Momentum resolution: 1% for pT > 1 GeV/c"]
- **Mitigation strategies**: [e.g., "Efficiency maps provided for tracking corrections"]
- **Run-specific conditions**: [e.g., subsystem downtime]
- …

6. Distribution / Access

- **Access:** [Internal repository, public release info]
- **File structure / Naming:** [e.g., ROOT DST files grouped by run segments]
- **License / Usage restrictions:** [Collaboration rules, redistribution policies]

7. Maintenance / Versioning

- **Version history:**
  - [v1.0 – initial release]
  - [v1.1 – updates if any]
- **Planned updates:** [Future improvements, corrections]
- **Contact for issues:** [Name, email]

8. Additional Notes

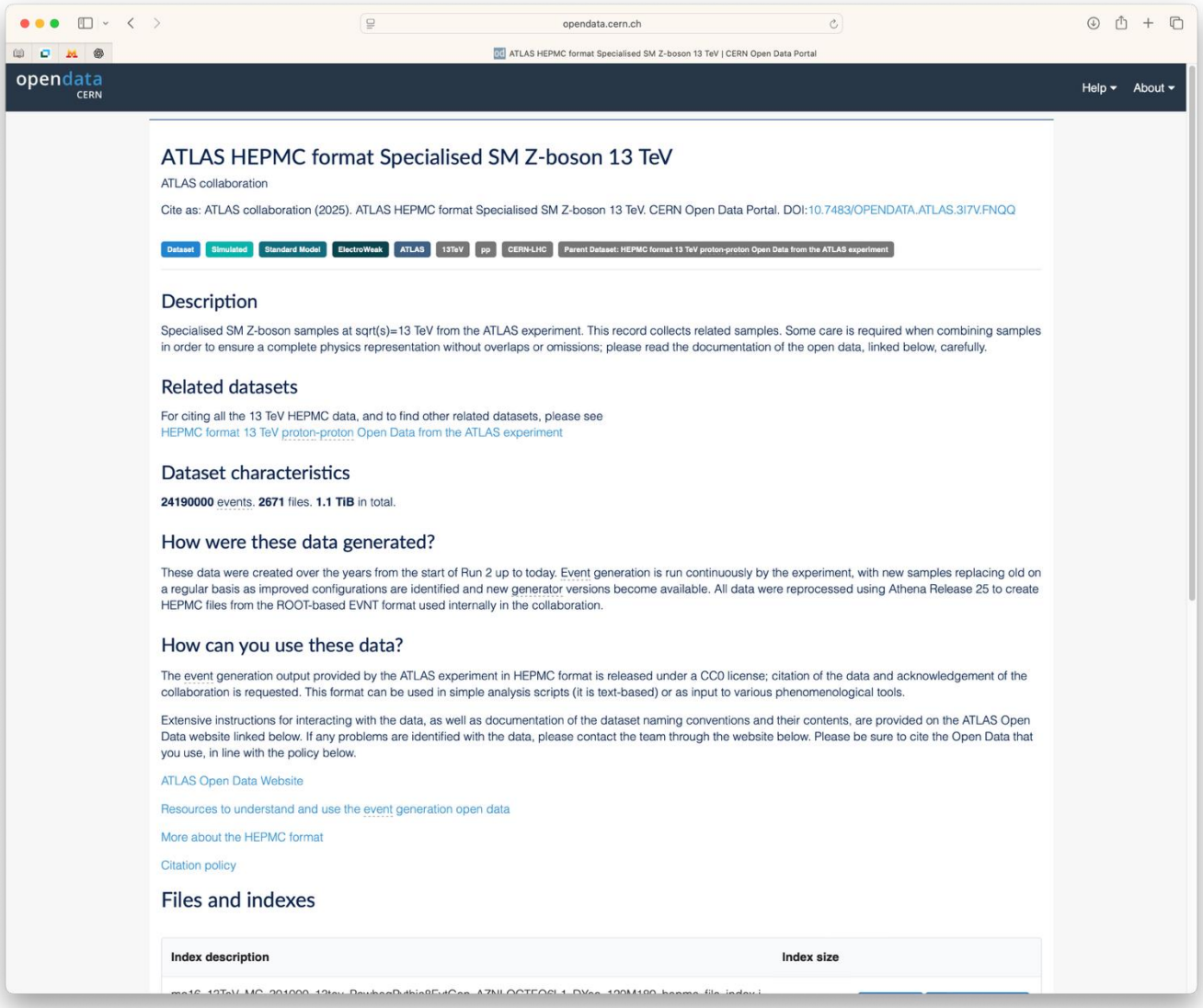
- **Related datasets**: [List any parent/child datasets or complementary datasets]
- **Publications**: [List DOIs or references for papers using this dataset]
- …

https://docs.google.com/document/d/176938hmHVBi2bO4144acbKeSAiOKId3nxGsLqeVeVCk/edit?usp=sharing

CERN Open Data

- Some elements of datasheets are accompanying open datasets
- Only open datasets!
- No standardization

https://opendata.cern.ch/record/160002



Why producing datasheets benefits RHIC?

- Preserves knowledge before it vanishes
  - Requires systematic documentation of expertise, assumptions, and caveats, some effort that pays on long term
- Creates a structured, clear, shared reference
  - Standardizes dataset descriptions across collaborations
  - Reduces misunderstandings and inconsistent analyses
- Enhances reproducibility and reusability
- Supports AI and automation