Datasheets for Nuclear Physics

What Are Datasheets?

Datasheets are structured documentation that provides the essential technical details needed to use a dataset

In the RHIC context, datasheets describe low-level data (NTuples/AODs) used for publications, and real or simulated data shared for technical work

Datasheets are Complementary to Existing Resources

Datasheets ≠ Papers ≠ HEPData

Journals: Provide scientific context, methodology, and results

HEPData: Store numerical results and figures

Datasheets: Add implementation details, software requirements, and implicit knowledge

Six Main Use Cases for Datasheets

Most relevant for Data and Analysis Preservation

Near-term Use Cases

1
Discovery & Assessment
Help users assess data suitability

2
Reproduction
Enable replication of published results

3
Development
Share technical data for algorithm work

Long-term Use Cases

4
Preservation
Maintain long-term usability

5
Education
Support training and learning

6
Citation
Ensure proper attribution

Use Case 1: Discovery & Assessment

Purpose:
Datasheets help researchers quickly determine if a dataset is suitable for their needs by providing rich metadata

Discovery is Metadata

Structured, searchable information enables researchers to find relevant datasets without downloading or analyzing the data itself

Example of key Metadata

- Collision system and energy
- Dataset size and statistics
- Kinematic coverage
- Available observables

Example:

Au+Au collisions at $\sqrt{s_{NN}}$ = 200 GeV from RHIC, 5M events

Similar to CERN Open Data portal entries that provide overview information

Beyond Metadata **Assessment** requires detailed technical documentation on its quality, limitations, and usability.

opendata.cern.ch

ATLAS ROOT ntuple format Run 2 2015+2016 proton-proton collision data beta release, 2J2LMET30 skim | CERN Open Data Portal

opendata CERN

Help About

ATLAS ROOT ntuple format Run 2 2015+2016 proton-proton collision data beta release, 2J2LMET30 skim

ATLAS collaboration

Cite as: ATLAS collaboration (2025). ATLAS ROOT ntuple format Run 2 2015+2016 proton-proton collision data beta release, 2J2LMET30 skim. CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.0CJR.N7ZT

Data recorded in 2015 and published in 2025

Dataset | Collision | ATLAS | 13TeV | pp | CERN-LHC | Parent Dataset: ROOT ntuple format 2015-2016 proton-proton Open Data for Education and Outreach from the ATLAS experiment

Metadata

Description

Run 2 2015+2016 proton-proton collision data beta release, 2J2LMET30 skim from the ATLAS experiment

Related datasets

For citing all the Open Data for Education and Outreach from this release, and to find other related datasets, please see
ROOT ntuple format 2015-2016 proton-proton Open Data for Education and Outreach beta release from the ATLAS experiment

Dataset characteristics

6242521 events. **16** files. **1.9 GiB** in total.

Assesment

How were these data selected?

These data were created during LS2 as part of a major reprocessing campaign of the Run 2 data. All data were reprocessed using Athena Release 22, and new corresponding MC simulation samples were produced. These data and MC simulation datasets were processed into ROOT ntuple files from the DAOD_PHYSLITE format that is released as open data for research. For the files in this record, the following skimming selection was applied: At least two jets with at least 20 GeV of p_T, at least two leptons passing tight identification requirements with at least 7 GeV of p_T, and 30 GeV of missing transverse momentum (i.e. a di-leptonic top-quark enhanced selection)

How can you use these data?

The data and MC simulation provided by the ATLAS experiment in root ntuple format is released under a CC0 license; citation of the data and acknowledgement of the collaboration is requested. This format can be used directly using ROOT or uproot for simple studies and is primarily intended for educational and outreach purposes.

Extensive instructions for interacting with the data, as well as documentation of the dataset naming conventions and their contents, are provided on the ATLAS Open Data website linked below. For those interested in implementing a research-quality data analysis, the open data designed for research (also linked below) may be a better starting point. Please be sure to cite the Open Data that you use, in line with the policy below.

ATLAS Open Data Website

Resources to understand and use the open data for education and outreach

More about this ntuple format

Ntuple making framework (PhysLiteToOpenData)

Citation policy

Files and indexes

Filename	Size

Use Case 2: Reproduction

Purpose:
Datasheets provide technical details essential for reproducing analysis results

Critical Information Often Missing from Papers

Software Environment

ROOT version, analysis framework, specific library versions

File Structure

Tree/branch names, variable definitions, storage conventions

Processing Steps

Quality cuts, calibrations, correction procedures

Example: ATLAS Open Data - https://opendata.cern.ch/record/80035
Dataset from Run 2 Pb-Pb collision data with detailed documentation but does not allow yet to redo an analysis

Datasheets complements publications by documenting implementation details

Use Case 4: Long-term Preservation

Purpose:
Datasheets maintain dataset usability for future researchers, even decades later

What Becomes Obsolete, Evolves or Gets Lost Over Time?

Software

Deprecated versions, discontinued libraries, changed APIs

ROOT 6, Ttree::Draw() API changes, Python 2.7

Conventions

Variable naming, unit systems, analysis frameworks

‘Centrality’, MeV/c vs GeV/c, variable ‘eta’

Context

Collaboration practices, detector status, running conditions

Good run list, correction files, documentation in obsolete wiki

Goes Beyond Reproduction
Supports new analyses using historical data

Essential for Legacy
Data from RHIC & other facility remain valuable long after operation ends

Without explicit documentation, datasets become harder to reuse as software environments change, community conventions evolve, and original tacit knowledge is lost

Use Case 5: Training & Education

Purpose:
Datasheets provide well-documented datasets for students and new researchers to learn analysis techniques

Educational Benefits

- Hands-on learning with real data
- Understanding analysis workflows
- Practicing data quality assessment
- Building reproducible analyses

Example: CERN Open Data

ALICE and other LHC experiments provide well-documented datasets with tutorials and getting-started guides that teach fundamental nuclear/particle physics analysis techniques

Requires clear, newcomer-friendly documentation that doesn't assume expert knowledge

Use Case 6: Compliance & Citation

Purpose

Datasheets provide formal, citable objects that ensure proper attribution and meet data sharing mandates

Three Functions

1

Credit

Acknowledge data producers when datasets are reused

2

Provenance

Track data lineage and authorship

3

Compliance

Meet agency requirements for data sharing

Datasheets themselves should carry DOIs, enabling proper citation.

Standardization

Standardized datasheets across nuclear physics will create a common language for data sharing

Community Benefits

- Enables cross-experiment comparisons
- Establishes best practices
- Supports tool interoperability

Sustainability

- Reduces burden on analysts and curators
- Consistent documentation across experiments
- Supports FAIR principles and long-term sustainability

Datasheets can be produced automatically by AI tools having access to internal information

Datasheets as AI-Ready Infrastructure

AI-Ready: Structured, machine-readable documentation that AI systems can parse, understand, and use to assist researchers with data analysis tasks

AI-Ready Documentation Enables

- Retrieval-Augmented Generation (RAG)
- Scientific chatbots and assistants
- Automated analysis suggestions
- Intelligent code generation

AI Systems Can

- Interpret experiment-specific formats
- Answer technical questions accurately
- Guide users through analysis workflows
- Suggest appropriate corrections/calibrations

Impact

Facilitates data exchange and reuse, and enables experiment-specific AI guidance

Learning from Existing Examples

CERN Open Data: Valuable but Incomplete

CERN's portal provides excellent starting points but highlights gaps that datasheets fill

What's Documented

- Dataset characteristics
- Selection criteria
- DOI and citation
- Documentation links

What's Often Missing

- Specific software versions
- Detailed variable definitions
- Quality flags meanings
- Known limitations
- Calibration details
- Reproduction workflows

Example: ATLAS record 93934 provides overview but requires navigation to multiple sites for implementation details opendata.cern.ch/record/93934

Standardized datasheets bridge these gaps

opendata.cern.ch

ATLAS DAOD_HION14 format Run 2 2015 Pb-Pb collision data | CERN Open Data Portal

opendata CERN

Help ▾ About ▾

ATLAS DAOD_HION14 format Run 2 2015 Pb-Pb collision data

ATLAS collaboration

Cite as: ATLAS collaboration (2024). ATLAS DAOD_HION14 format Run 2 2015 Pb-Pb collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.GUZD.H9UF

Data recorded in 2015 and published in 2024

Dataset | Collision | Heavy-ion physics | ATLAS | 5TeV | PbPb | CERN-LHC | Parent Dataset: DAOD_HION14 format 2015 Pb-Pb Open Data for Research from the ATLAS experiment

Description

Run 2 2015 Pb-Pb collision data from the ATLAS experiment

Related datasets

For citing all the Heavy Ion Open Data for Research from this release, and to find other related datasets, please see
DAOD_HION14 format 2015 Pb-Pb Open Data for Research from the ATLAS experiment

Dataset characteristics

220950858 events. **1913** files. **4.0 TiB** in total.

How were these data selected?

These data were created during LS2 as part of a major reprocessing campaign of the Run 2 data. All data were reprocessed using Athena Release 21, and new corresponding MC simulation samples were produced. These data and MC simulation datasets were processed into DAOD_HION14 format files; this is a light-weight data format intended for analysis of heavy ion minimum bias data, sufficient to support a wide variety of ATLAS analyses.

How can you use these data?

The data and MC simulation provided by the ATLAS experiment in DAOD_HION14 format is released under a CC0 license; citation of the data and acknowledgement of the collaboration is requested. This format can be used directly like a ROOT ntuple (or using uproot) for simple studies or processed into secondary ntuples with systematic uncertainties included using the ATLAS AnalysisBase software.

Extensive instructions for interacting with the data, as well as documentation of the dataset naming conventions and their contents, are provided on the ATLAS Open Data website linked below. Designing and implementing a research-quality data analysis is a complex process that requires an understanding of particle physics; for those new to the subject, the open data designed for education (also linked below) might be a good starting point. Please be sure to cite the Open Data that you use, in line with the policy below.

ATLAS Open Data Website

Resources to understand and use the open data for research

ATLAS Analysis Software Tutorial

More about the DAOD_HION14 data format

Citation policy

Files and indexes

Index description	Index size

Proposal: RHIC Data Datasheets

Goal:
Create standardized datasheets for RHIC datasets used in publications to enable re-analysis and long-term preservation

Essential Information to Include

Dataset Specifications

- Collision system, energy, year
- Event selection criteria
- Trigger requirements
- Centrality definitions

Technical Details

- ROOT version and dependencies
- Tree/branch structure
- Variable units and conventions
- Calibration/correction status

Quality Information

- Data quality flags
- Known issues or limitations
- Recommended usage
- Contact information

Create standardized, machine-readable datasheets for RHIC datasets to support reuse, reproducibility, and preservation.

Format: Machine-readable with associated DOI

Example

Measurement of Sequential Υ Suppression in Au + Au Collisions at sNN = 200 GeV with the STAR Experiment

Phys. Rev. Lett. **130**, 112301 – **Published 14 March, 2023**
DOI: https://doi.org/10.1103/PhysRevLett.130.112301


PHYSICAL REVIEW LETTERS **130**, 112301 (2023)

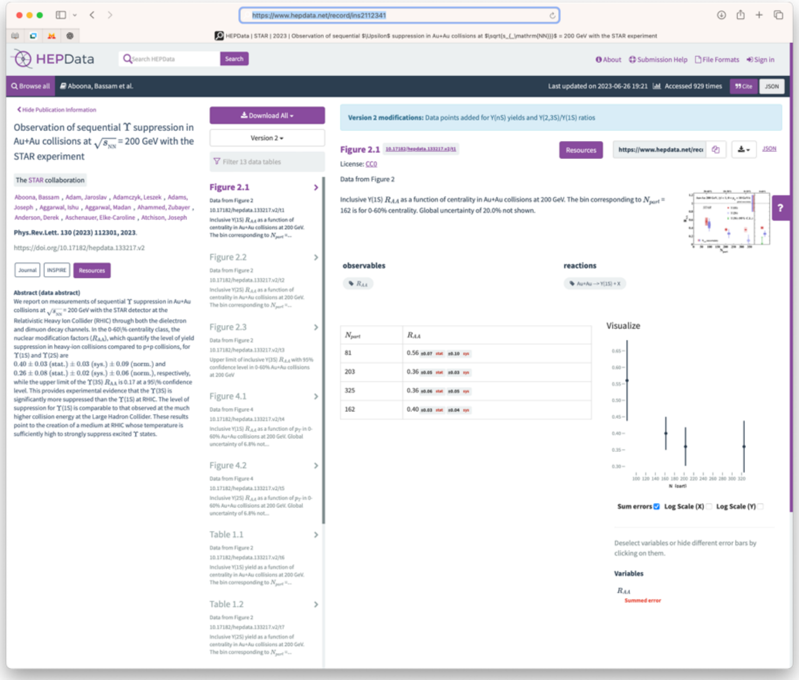
Measurement of Sequential Υ Suppression in Au + Au Collisions at $\sqrt{s_{NN}} = 200$ GeV with the STAR Experiment

(STAR Collaboration)

[1]*Abilene Christian University, Abilene, Texas 79699*
[2]*AGH University of Science and Technology, FPACS, Cracow 30-059, Poland*

0031-9007/23/130(11)/112301(8) 112301-1 © 2023 American Physical Society

https://www.hepdata.net/record/ins2112341



- **HEPData:** "*Here are the published numbers for reuse/citation*"
- **Datasheets:** "*Here's how to work with the actual dataset files*"

Building a Complete Datasheet

From the Paper

Physics Context & Methods

- Collision system: Au+Au at $\sqrt{s_{NN}}$ = 200 GeV
- Detectors: TPC, BEMC, MTD
- Triggers: BEMC ET > 3.5 GeV, MTD dimuon
- Centrality: 0-10%, 10-30%, 30-60%
- Kinematic coverage: $|y| < 1$, $p_T < 10$ GeV/c

Analysis Approach

- Unbinned ML fit (mentioned, not detailed)
- Like-sign background subtraction
- GEANT3 simulations (version unspecified)
- $b\bar{b}$/Drell-Yan from PYTHIA6

Qualitative Results

- R_{AA} suppression patterns vs centrality
- R_{AA} vs p_T trends
- Model comparisons (OQS+pNRQCD, transport)
- Comparison to LHC (CMS 5.02 TeV)

From HEPData

Numerical Results (Machine-Readable)

- R_{AA} values for Υ(1S), Υ(2S), Υ(3S) upper limit
- Statistical uncertainties
- Systematic uncertainties (by source)
- N_{part} and p_T bin edges
- Tables in CSV, YAML, ROOT, JSON

Metadata

- DOI for HEPData dataset
- CC-BY 4.0 license
- Version tracking
- Links to paper (DOI, arXiv, INSPIRE)
- Searchable by system/observable/energy

What HEPData Provides

✓ Published results for citation/reuse
✓ Exact numerical values from figures
✓ Standardized formats for plotting
✗ Not the underlying dataset files
✗ Not how to reproduce from data

Gap
Neither the paper nor HEPData includes the technical details required to access, understand, or process the dataset.

What Datasheets Add

> Datasheets document the practical details needed to work with actual dataset files

Dataset Access & Format

- Location of ROOT files
- Tree/branch structure and variable definitions
- Access instructions
- Data volume, file sizes
- Event/candidate counts per file

Selection & Reconstruction

- Exact cut values (dE/dx windows, E/p, DCA)
- Track quality criteria (nhits, χ^2/ndf)
- Trigger configuration details
- PID working points
- Momentum/energy correction procedures
- Vertex finding algorithms

Computational Reproducibility

- ROOT version, STAR framework version
- GEANT3 version and configuration
- Analysis code (ROOT macros, scripts)
- Workflow description (pipeline steps)
- Calibration file references

Usage & Quality

- Good run lists, bad run lists
- Data quality flags and meanings
- Known issues or limitations
- Example code snippets
- Recommended usage for different goals
- Contact information for questions

Datasheets document the **tacit knowledge** that otherwise disappears

Summary: Datasheets in Nuclear Physics

Datasheets capture implicit technical knowledge

Six use cases for Datasheets have been identified to support scientific needs

Standardization enables sustainable AI-ready nuclear physics datasets

Next step: Prototype with recent publication?