

# **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  $\neq$  Papers  $\neq$  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

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

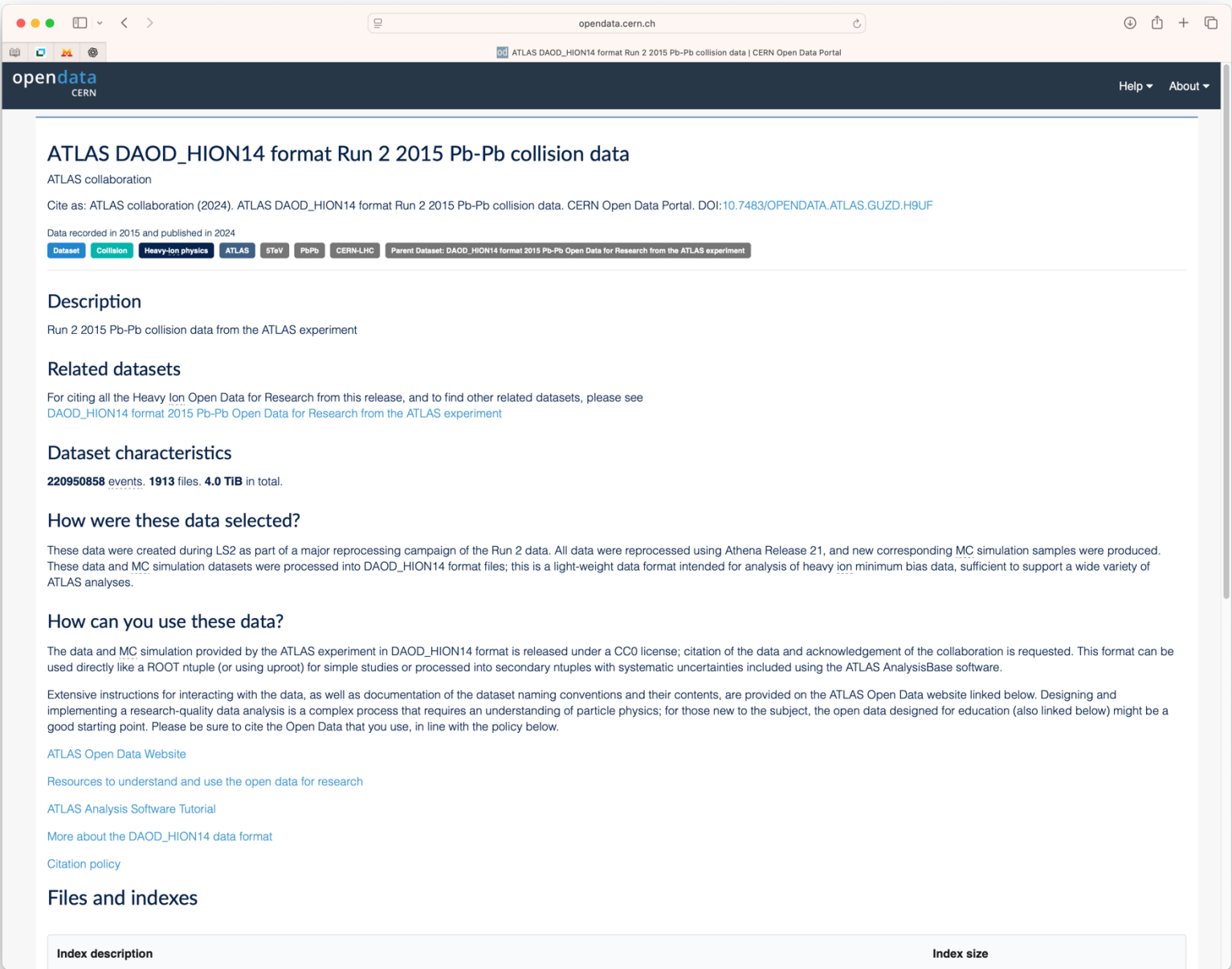
- Dataset characteristics
- Selection criteria
- DOI and citation
- Documentation links

### 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](https://opendata.cern.ch/record/93934)

**Standardized datasheets bridge these gaps**



# 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 $\Upsilon$ Suppression in Au + Au Collisions at $\sqrt{s_{NN}} = 200$ GeV with the STAR Experiment

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

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

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

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

B. E. Aboona,<sup>53</sup> J. Adam,<sup>15</sup> L. Adamczyk,<sup>2</sup> J. R. Adams,<sup>38</sup> I. Aggarwal,<sup>40</sup> M. M. Aggarwal,<sup>40</sup> Z. Ahammed,<sup>59</sup> D. M. Anderson,<sup>53</sup> E. C. Aschenauer,<sup>2</sup> J. Alchison,<sup>1</sup> V. Bairathi,<sup>21</sup> W. Baker,<sup>1</sup> J. G. Ball Cup,<sup>1</sup> K. Barish,<sup>1</sup> R. Bellwied,<sup>21</sup> P. Bhattacharya,<sup>28</sup> A. Bhasin,<sup>28</sup> S. Bhatta,<sup>28</sup> J. Bielcik,<sup>15</sup> J. Bielcikova,<sup>15</sup> J. D. Brandenburg,<sup>15</sup> X. Z. Cai,<sup>54</sup> H. Caines,<sup>45</sup> M. Calderón de la Barca Sánchez,<sup>2</sup> D. Cebra,<sup>1</sup> J. Ceska,<sup>15</sup> I. Chakaberia,<sup>21</sup> P. Chaloupka,<sup>15</sup> B. K. Chan,<sup>10</sup> Z. Chang,<sup>26</sup> D. Chen,<sup>11</sup> J. Chen,<sup>47</sup> J. H. Chen,<sup>19</sup> Z. Chen,<sup>47</sup> J. Cheng,<sup>55</sup> Y. Cheng,<sup>10</sup> S. Choudhury,<sup>19</sup> W. Christie,<sup>6</sup> X. Chu,<sup>6</sup> H. J. Crawford,<sup>6</sup> M. Csanád,<sup>17</sup> G. Dale-Gau,<sup>13</sup> A. Das,<sup>15</sup> M. Daugherty,<sup>21</sup> J. M. Deppner,<sup>20</sup> A. Dhamija,<sup>40</sup> L. Di Carlo,<sup>61</sup> L. Didenko,<sup>6</sup> P. Dixit,<sup>23</sup> X. Dong,<sup>1</sup> J. L. Drachenberg,<sup>1</sup> E. Duckworth,<sup>21</sup> J. C. Dunlop,<sup>6</sup> J. Engelage,<sup>6</sup> G. Eppley,<sup>42</sup> S. Esumi,<sup>56</sup> O. Evdokimov,<sup>15</sup> A. Ewigleben,<sup>20</sup> O. Eyer,<sup>6</sup> R. Fatemi,<sup>20</sup> S. Fazio,<sup>1</sup> C. J. Feng,<sup>36</sup> Y. Feng,<sup>41</sup> E. Finch,<sup>49</sup> Y. Fisyak,<sup>6</sup> F. A. Flor,<sup>12</sup> C. A. Gagliardi,<sup>21</sup> T. Galatyuk,<sup>16</sup> F. Geurts,<sup>2</sup> N. Ghimire,<sup>2</sup> A. Gibson,<sup>28</sup> K. Gopali,<sup>24</sup> X. Guo,<sup>1</sup> D. Grosnick,<sup>28</sup> A. Gupta,<sup>2</sup> W. Gyun,<sup>5</sup> A. Hamed,<sup>2</sup> Y. Han,<sup>42</sup> S. Harabasz,<sup>6</sup> M. D. Harasty,<sup>1</sup> J. W. Harris,<sup>42</sup> H. Harrison,<sup>28</sup> W. He,<sup>2</sup> X. H. He,<sup>2</sup> Y. He,<sup>1</sup> S. Heppelmann,<sup>2</sup> N. Hermann,<sup>2</sup> L. Holuh,<sup>6</sup> C. Hu,<sup>2</sup> Q. Hu,<sup>2</sup> Y. Hu,<sup>21</sup> H. Huang,<sup>28</sup> H. Z. Huang,<sup>28</sup> G. L. Huang,<sup>15</sup> Y. Huang,<sup>2</sup> Y. Huang,<sup>2</sup> T. J. Hummer,<sup>38</sup> D. Isenhardt,<sup>2</sup> M. Iosifaki,<sup>28</sup> W. W. Jacobs,<sup>28</sup> A. Jaloira,<sup>28</sup> C. Jena,<sup>28</sup> A. Jentsch,<sup>2</sup> Y. Ji,<sup>31</sup> J. Jia,<sup>6,20</sup> C. Jin,<sup>42</sup> X. Ju,<sup>45</sup> E. G. Judd,<sup>6</sup> S. Kahana,<sup>51</sup> M. L. Kahir,<sup>11</sup> S. Kaganaster,<sup>32</sup> D. Kalinkin,<sup>30,6</sup> K. Kang,<sup>35</sup> D. Kapukchyan,<sup>11</sup> K. Kauder,<sup>6</sup> H. W. Ke,<sup>2</sup> D. Keane,<sup>29</sup> M. Kelsey,<sup>41</sup> Y. V. Khaynzhi,<sup>38</sup> D. P. Kikola,<sup>40</sup> B. Kimelman,<sup>2</sup> D. Kinsies,<sup>1</sup> I. Kisel,<sup>18</sup> A. Kiselev,<sup>6</sup> A. G. Klocsey,<sup>32</sup> H. S. Ko,<sup>1</sup> L. K. Kosarzewski,<sup>15</sup> L. Kramarik,<sup>15</sup> L. Kumar,<sup>40</sup> S. Kumar,<sup>27</sup> R. Kunnawalkam Elayavalli,<sup>62</sup> R. R. Lacey,<sup>20</sup> J. M. Landgraf,<sup>1</sup> J. Lauret,<sup>6</sup> A. Lebedev,<sup>1</sup> J. H. Lee,<sup>6</sup> Y. H. Leung,<sup>20</sup> N. Lewis,<sup>6</sup> C. Li,<sup>47</sup> C. Li,<sup>45</sup> W. Li,<sup>45</sup> X. Li,<sup>45</sup> Y. Li,<sup>45</sup> Y. Li,<sup>45</sup> Z. Li,<sup>45</sup> X. Liang,<sup>11</sup> Y. Liang,<sup>28</sup> R. Licanik,<sup>17,15</sup> T. Lin,<sup>6</sup> M. A. Lisa,<sup>28</sup> C. Liu,<sup>2</sup> F. Liu,<sup>12</sup> H. Liu,<sup>28</sup> H. Liu,<sup>12</sup> L. Liu,<sup>12</sup> T. Liu,<sup>62</sup> X. Liu,<sup>62</sup> Y. Liu,<sup>2</sup> Z. Liu,<sup>2</sup> T. Ljubicic,<sup>6</sup> W. J. Llope,<sup>40</sup> O. Lomicky,<sup>15</sup> R. S. Longacre,<sup>6</sup> E. Loyd,<sup>1</sup> T. Lu,<sup>27</sup> N. S. Lukow,<sup>25</sup> X. F. Luo,<sup>1</sup> L. Ma,<sup>1</sup> R. Ma,<sup>2</sup> Y. G. Ma,<sup>1</sup> N. Magdy,<sup>30</sup> D. Mallick,<sup>15</sup> S. Margetis,<sup>2</sup> C. Markert,<sup>28</sup> H. S. Mats,<sup>1</sup> J. A. Mazer,<sup>41</sup> G. McNamara,<sup>41</sup> K. Mi,<sup>15</sup> S. Mioduszewski,<sup>15</sup> B. Mohanty,<sup>21</sup> I. Mooney,<sup>20</sup> A. Mukherjee,<sup>17</sup> M. I. Nagy,<sup>17</sup> A. S. Nain,<sup>40</sup> J. D. Nam,<sup>28</sup> Md. Nasim,<sup>21</sup> D. Neff,<sup>10</sup> J. M. Nelson,<sup>42</sup> D. B. Nemes,<sup>62</sup> M. Nie,<sup>42</sup> T. Niida,<sup>26</sup> R. Nishitani,<sup>26</sup> T. Nonaka,<sup>28</sup> A. S. Nunes,<sup>6</sup> G. Odyniec,<sup>1</sup> A. Ogawa,<sup>2</sup> S. Oh,<sup>31</sup> K. Okubo,<sup>26</sup> B. S. Page,<sup>6</sup> R. Pak,<sup>1</sup> J. Pan,<sup>53</sup> A. Pandey,<sup>35</sup> A. K. Pandey,<sup>27</sup> T. Pani,<sup>43</sup> A. Paul,<sup>11</sup> B. Pawlik,<sup>20</sup> D. Pawlowska,<sup>40</sup> C. Perkins,<sup>4</sup> J. Pluta,<sup>40</sup> B. R. Pokhrel,<sup>32</sup> M. Posik,<sup>32</sup> T. Proizman,<sup>32</sup> V. Prozorova,<sup>15</sup> N. K. Pruthi,<sup>40</sup> M. Przybycien,<sup>2</sup> J. Putschke,<sup>1</sup> Z. Qin,<sup>55</sup> H. Qiu,<sup>27</sup> A. Quintero,<sup>52</sup> C. Racz,<sup>15</sup> S. K. Radhakrishnan,<sup>29</sup> N. Raha,<sup>61</sup> R. L. Ray,<sup>54</sup> R. Reed,<sup>15</sup> H. G. Ritter,<sup>31</sup> C. W. Robertson,<sup>41</sup> M. Robotkova,<sup>17,15</sup> J. L. Romero,<sup>6</sup> M. A. Rosales Aguilar,<sup>40</sup> D. Roy,<sup>11</sup> P. Roy Chowdhury,<sup>40</sup> L. Ruan,<sup>4</sup> A. K. Sahoo,<sup>21</sup> N. R. Sahoo,<sup>47</sup> H. Sako,<sup>56</sup> S. Salur,<sup>2</sup> S. Sato,<sup>28</sup> W. B. Schmidtke,<sup>2</sup> N. Schmitz,<sup>19</sup> F. J. Seck,<sup>19</sup> J. Seger,<sup>19</sup> R. Seto,<sup>1</sup> P. Seyboth,<sup>19</sup> N. Shah,<sup>29</sup> P. V. Sharmuganathan,<sup>1</sup> M. Shao,<sup>1</sup> T. Shao,<sup>1</sup> M. Sharma,<sup>2</sup> N. Sharma,<sup>2</sup> R. Sharma,<sup>2</sup> S. R. Sharma,<sup>2</sup> A. I. Sheikh,<sup>29</sup> D. Y. Shen,<sup>2</sup> K. Shen,<sup>45</sup> S. S. Shi,<sup>12</sup> Y. Shi,<sup>12</sup> Q. Y. Shou,<sup>19</sup> F. Si,<sup>41</sup> J. Singh,<sup>28</sup> S. Singha,<sup>27</sup> P. Sinha,<sup>24</sup> M. J. Skoby,<sup>3,41</sup> N. Smirnov,<sup>62</sup> Y. Söngen,<sup>20</sup> Y. Song,<sup>12</sup> B. Srivastava,<sup>41</sup> T. D. S. Stanislaus,<sup>2</sup> M. Stefania,<sup>2</sup> D. J. Stewart,<sup>41</sup> B. Stringfellow,<sup>41</sup> Y. Su,<sup>45</sup> A. A. P. Suaide,<sup>14</sup> M. Sumera,<sup>37</sup> C. Sun,<sup>30</sup> X. Sun,<sup>27</sup> Y. Sun,<sup>45</sup> Y. Sun,<sup>27</sup> B. Surrow,<sup>42</sup> Z. W. Sweger,<sup>2</sup> P. Szymanski,<sup>40</sup> A. Tami,<sup>42</sup> A. H. Tang,<sup>42</sup> T. Tang,<sup>42</sup> T. Tarnowsky,<sup>34</sup> J. H. Thomas,<sup>31</sup> A. R. Timmins,<sup>21</sup> D. Tlustý,<sup>14</sup> T. Todoroki,<sup>36</sup> C. A. Tomkiet,<sup>32</sup> S. Trentalange,<sup>10</sup> R. E. Tribble,<sup>53</sup> P. Tribedy,<sup>6</sup> T. Truhlar,<sup>15</sup> B. A. Trzeciak,<sup>15</sup> O. D. Tsai,<sup>10,6</sup> C. Y. Tsang,<sup>26,6</sup> Z. Tu,<sup>1</sup> T. Ullrich,<sup>40</sup> D. G. Underwood,<sup>15,18</sup> L. Upsal,<sup>42</sup> G. Van Buren,<sup>6</sup> J. Vaneck,<sup>4</sup> I. Vassiliev,<sup>18</sup> V. Verkeste,<sup>21</sup> F. Videbæk,<sup>6</sup> S. A. Voloshin,<sup>1</sup> F. Wang,<sup>40</sup> G. Wang,<sup>40</sup> J. S. Wang,<sup>27</sup> X. Wang,<sup>47</sup> Y. Wang,<sup>47</sup> Y. Wang,<sup>12</sup> Y. Wang,<sup>27</sup> Z. Wang,<sup>11</sup> J. C. Webb,<sup>4</sup> P. C. Weidenkaff,<sup>10</sup> G. D. Westfall,<sup>10</sup> D. Wielanek,<sup>40</sup> H. Wieman,<sup>41</sup> G. Wilks,<sup>31</sup> S. W. Wissink,<sup>28</sup> R. Witt,<sup>1</sup> J. Wu,<sup>1</sup> J. Wu,<sup>1</sup> X. Wu,<sup>1</sup> Y. Wu,<sup>1</sup> B. Xi,<sup>1</sup> Z. G. Xiao,<sup>1</sup> W. Xie,<sup>1</sup> H. Xu,<sup>21</sup> N. Xu,<sup>42</sup> Q. H. Xu,<sup>47</sup> Y. Xu,<sup>47</sup> Y. Xu,<sup>47</sup> Z. Xu,<sup>10</sup> G. Yan,<sup>47</sup> C. Yang,<sup>47</sup> Q. Yang,<sup>47</sup> S. Yang,<sup>46</sup> Y. Yang,<sup>20</sup> Z. Ye,<sup>42</sup> Z. Ye,<sup>11</sup> L. Yi,<sup>41</sup> C. K. Yip,<sup>4</sup> Y. Y. Y. Zhong,<sup>60</sup> W. Zhao,<sup>45</sup> C. Zhang,<sup>12</sup> J. Zhang,<sup>12</sup> J. Zhang,<sup>12</sup> S. Zhang,<sup>45</sup> X. Zhang,<sup>27</sup> Y. Zhang,<sup>27</sup> Y. Zhang,<sup>28</sup> Y. Zhang,<sup>12</sup> J. Zhang,<sup>36</sup> Z. Zhang,<sup>45</sup> Z. Zhang,<sup>12</sup> F. Zhao,<sup>27</sup> J. Zhao,<sup>19</sup> M. Zhao,<sup>6</sup> C. Zhou,<sup>19</sup> J. Zhou,<sup>45</sup> S. Zhou,<sup>12</sup> Y. Zhou,<sup>12</sup> X. Zhu,<sup>12</sup> S. M. Zurek,<sup>2</sup> and M. Zyzak,<sup>18</sup>

(STAR Collaboration)

<sup>1</sup>Ahline Christian University, Ahline, Texas 79659  
<sup>2</sup>AGH University of Science and Technology, PPACS, Cracow 30-059, Poland

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

HEPData (STAR) (2023) Observation of sequential  $\Upsilon$  suppression in Au+Au collisions at  $\sqrt{s_{NN}} = 200$  GeV with the STAR experiment

Version 2 modifications: Data points added for  $\Upsilon$ 1S1 yields and  $\Upsilon$ 2S1/ $\Upsilon$ 1S1 ratios

Figure 2.1 Inclusive  $\Upsilon$ 1S1  $R_{AA}$  as a function of centrality in Au+Au collisions at 200 GeV. The bin corresponding to  $N_{part} = 152$  is for 0-40% centrality. Global uncertainty of 20-25% not shown.

Centrality	$R_{AA}$	$N_{part}$
0-5%	0.56 ± 0.07	81
5-10%	0.56 ± 0.07	101
10-20%	0.56 ± 0.07	152
20-30%	0.56 ± 0.07	201
30-40%	0.56 ± 0.07	251
40-50%	0.56 ± 0.07	301
50-60%	0.56 ± 0.07	351
60-70%	0.56 ± 0.07	401
70-80%	0.56 ± 0.07	451
80-90%	0.56 ± 0.07	501
90-100%	0.56 ± 0.07	551

- 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)
- bb̄/Drell-Yan from PYTHIA6

### Qualitative Results

- RAA suppression patterns vs centrality
- RAA vs  $p_T$  trends
- Model comparisons (OQS+pNRQCD, transport)
- Comparison to LHC (CMS 5.02 TeV)

## From HEPData

### Numerical Results (Machine-Readable)

- RAA values for Y(1S), Y(2S), Y(3S) upper limit
- Statistical uncertainties
- Systematic uncertainties (by source)
- Npart 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,  $\chi^2/\text{ndf}$ )
- 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?