Follow-up on Metadata for DAP

Review report & Metadata

(Meta)data curation must begin as early as possible within the experiments themselves. This is a critical prerequisite for the success of the project and cannot be outsourced or delayed, as well-structured metadata is essential for both reproducibility and effective AI integration.

External review emphasized early metadata capture as critical to RHIC data usability.

Data Preservation: Metadata Is Not Optional

Without Context, Data Becomes Useless

- Experiments have generated terabytes of data, but within 5–10 years, critical context will likely disappear (there a plenty of example even at RHIC).
- Even when files remain accessible, they become unusable without proper metadata.
- Data without context or metadata rapidly loses any scientific value.

Without Metadata	With Metadata
Software version unknown → results not reproducible	Software v2.3.1, config file reco_2018.yaml
Beam energy or magnet polarity unknown.	Run configuration and beam energy.
Calibration lost → data quality uncertain	Calibration snapshot linked to dataset
Generator settings missing → simulation unusable	$\sqrt{s_{nn}}$ = 200 GeV, GEANT3.x, PHENIX software

Metadata ensures that RHIC's archived data remain scientifically usable long after operations end.

If someone needs to ask:

"how did you make this?",
"how do I use this?", or
"how do I find datasets like this?",

...then that information should be in the **metadata**.

Metadata Enables Future Reserach

- Without metadata, future researchers or users of RHIC data cannot locate or trust datasets.
- With metadata, future users can search for *“Au+Au 200 GeV minimum-bias runs with TPC and TOF data (2014)”* and find them.
- Missing metadata wastes unique RHIC data that can never be re-collected.

Metadata is essential for findability, validation, and reproducibility.

What Metadata to Capture

Metadata answers three questions:

What is it? How was it made? How can it be used?

Three classes of Metadata

1/ ***Content****, 2/* ***Processing****, 3/* ***Context***

Content Metadata: *What is this dataset?*

- Au+Au collisions, Run 14, 200 GeV, integrated luminosity = 1.2 nb^{-1}
- sPHENIX simulation, 10 M events, ROOT files
- DOI: 10.5281/zenodo.12345

Processing Metadata: *How was it made?*

- STAR software release SL24c, commit 7e9bcd2
- GEANT 4.10, HIJING 1.411, tune 2023-A
- Steps: raw → DST → ntuples

Context Metadata: *How to use it?*

- Requires: runlog 2024.xml, gain_map_v5.root
- Tools: ROOT 6.30+, Python 3.11
- Known issue: TPC sector 17 high-voltage instability
- Reference: STAR Collab., *Phys. Rev. C 108 (2024) 045203*

Metadata is the bridge between RHIC's preserved files and future scientific use

Metadata for future users

- We need to prototype long term solutions that will allow **future users** to make sense of the data that are stored.
- Not all existing information needs to be preserved
- Will existing solutions used by experiments be adequate, understandable and accessible for future users?
- Will current solutions be maintainable long term?
- We need to investigate and prototype