DRAFT - RHIC DAP Round Table #4

02/04/2025

https://indico.bnl.gov/event/26327/

1. Introduction & Standards for Repositories

- The CoreTrustSeal model for data repositories provides a best practices framework for ensuring that data is preserved in a sustainable and accessible manner.
- The current dependence on HPSS (High-Performance Storage System) for data storage might not be sustainable in the long run. Participants discussed potential alternatives, including distributed storage systems.
- Collaboration and knowledge sharing within the community are essential for addressing data preservation challenges. Participants discussed possible partnerships with other institutions and the opportunity to learn from successful data preservation initiatives at CERN.

2. PHENIX Answers to the Questionnaire

- **Data Volume, Organization & Storage:**
  - The PHENIX RAW data volume is about 23PB. There are also a few intermediate dataset DSTs (Volume?)
  - The data are mostly stored on HPSS, some on GPFS (Volume?), and recently, dCache. (Volume?)
  - There is no currently working data catalog, data are identified through their name. Data can be discovered through a ROOT macro stored on CVS.
  - Datasets used for final analysis are usually stored in some user areas (though they are required to be stored in HPSS). HEPData is used to archive final data points and tables.
- **Data Management**
  - Two main data formats are used, PRDF (PHENIX Raw Data Files) and ROOT for DTSs.
  - Going back to RAW data is unlikely, meaning that DAP Level 3 is the target.

    - DSTs can be used for ‘recalibration ‘ through the ‘analysis taxi’
    - Typical published analysis should be accompanied by a list of runs (indicated in analysis note)
- **Metadata**
    - Metadata are stored in the file name, eg, Verbal, descriptive names for datasets, for instance *Run15pp200CAERTP108* - Central Arm, ERT triggered data, Production 108 (reference to code in CVS)
- **Conditions Data**
    - Two Databases are mainly used: the run database and the calibration database.
    - Volume? Where are they stored? Some are on CVS (all?)
- **Software**
    - The software used for the analysis is archived (zip) in HPSS. Time of the creation of the analysis note.
    - Analysis software is in C++. Analysis code is on AFS and CVS
- Workflows
    - Apart from the taxi system, analysis is performed manually
- **Preservation – Documentation**
    - An un-initiated person cannot reproduce the majority of analyses. One analysis has been preserved and reproduced by an external person.
    - A new web site based on Jekyll, has been designed to store documentation related to preservation.
    - Results are stored on HEPData
    - Zenodo at CERN has been used to store notes and documentation
- **Data Sharing**
    - A small set of calorimeter data is available on the OpenData portal, illustrating the photon and pi0 analysis.
    - Data stored at BNL requires a BNL account to access them.

Other discussions:

- A discussion also mentions the ALEPH experiment at LEP and the H1 experiment at DESY. In 2024, the data from these experiments was reanalyzed using algorithms developed long after data collection ended. Can we learn from them?

- Jerome expressed concerns regarding the long-term viability of containerization for software preservation because of the rapid evolution of technology.
- Christine suggested including data from other experiments, such as Phobos and Brahms, in the preservation efforts.

3. Follow up

- Provide data volumes for PHENIX DSTs and other formats on disk.
- Details about PHENIX databases.
- Clarify whether the PHENIX taxi system needs to be maintained to preserve its analysis capability and what keeping it would entail.
- List of the PHENIX external dependencies (e.g., ROOT, HPSS, CVS, AFS, GitHub, Docker, etc.) required for Level 3.
- Contact Yen-Jie Lee at MIT for insights about reanalyzing ALEPH data in particular.