RHIC Data Analysis and Preservation

Status Report and Implementation Plan

This report summarizes data and analysis preservation activities since the last PAC.

In November 2024, the PAC requested *"a detailed report on the status and plans for long-term RHIC data analysis and data preservation, including workforce, computing resources, and possible timelines."*

In response, BNL developed a comprehensive RHIC Data Analysis and Preservation (DAP) Plan. This plan is structured into two phases and has been externally reviewed. Phase I (2026–2030) focuses on establishing and integrating tools and methods while collaborations and infrastructure remain active. Phase II (2031 and beyond) transitions to sustainable stewardship with fewer resources, while still maintaining the essential capabilities for long-term preservation.

The DAP Plan builds upon best practices from the Data Preservation in High Energy Physics (DPHEP) collaboration and the ICFA Data Lifecycle Panel. It leverages expertise from CERN and DESY and aligns with FAIR principles (Findable, Accessible, Interoperable, Reusable).

During Phase I, the DAP effort will be organized by a core team focusing on software preservation, repositories, AI integration, documentation, and collaboration with experiment liaisons. The current hardware is adequate for Phase I, but a replacement will be necessary around 2031 as storage and computing systems reach end-of-life. At the end of Phase I, a global reprocessing campaign of RHIC data might be conducted to produce consistent, fully documented datasets using final calibrations and software.

Preservation Strategy and Approach

Preserving RHIC data is challenging because of its large scale, hundreds of petabytes, and the variety of experimental conditions accumulated over two decades. Multiple collision systems, beam energies, and detector setups demand thorough documentation and well-maintained analysis environments.

The DAP Plan ensures continued access to processed physics data, analysis software, and documentation, while archiving all raw data. Large-scale reprocessing is not planned due to the necessary computing resources, but it remains a potential future option. Analysis environments are preserved through containerization, capturing complete software stacks and workflows as reproducible packages. This approach guarantees that analyses remain executable and understandable long after the original developers have moved on.

Recent Developments

Since the last PAC meeting, significant progress has been made. Regular DAP roundtables now bring together representatives from PHENIX, sPHENIX, and STAR, along with external experts, to coordinate activities and share expertise.

A significant innovation is a ChatBot that uses large language models combined with Retrieval-Augmented Generation. Thousands of documents, including technical notes, manuals, calibration procedures, and analysis guides, have been indexed and cataloged. Researchers can now ask complex questions in natural language and receive answers directly from this knowledge base, reducing dependence on multiple experts.

A multi-experiment InvenioRDM repository has been deployed at BNL to manage both public and restricted documents, incorporating version control and provenance tracking. It connects with OSTI for DOI assignment and supports federated authentication, allowing seamless cross-institutional access.

The CERN Open Data Portal has been adapted and implemented at BNL, initially populated with PHENIX datasets. It provides metadata-based search capabilities and enforces access restrictions depending on data sensitivity.

These advancements, including coordination forums, AI knowledge capture, repositories, and public data portals, establish the technical and organizational foundation for long-term preservation. They help reduce implementation risks and demonstrate that the DAP Plan approach is feasible.

External Review and Validation

An international committee of data preservation experts reviewed the DAP Plan in July 2025. The committee strongly supported the phased approach as cost-effective and technically sound, highlighting the use of containerized workflows and AI-driven knowledge navigation. They emphasized two key challenges: metadata curation must begin while expert knowledge is still available, and institutional commitments are vital since collaborations do not control personnel allocations. The committee also approved the DAP core team workforce plan and proposed governance.

Workforce and Timeline

Phase I staffing supports prototyping, workflow development, and knowledge capture while collaborations remain fully engaged. Phase II staffing ensures ongoing stewardship with fewer resources. A governance structure with experiment representatives, BNL management, and external advisors provides oversight and long-term continuity.

- Phase I (2026–2030): 6.5 FTE core team, with 3–5 researchers per experiment contributing 15–25% effort. Key deliverables include metadata curation, containerized workflows, AI integration, and a planned global reprocessing campaign.

- Phase II (2031+): 2.1 FTE liaisons support sustainable operations, periodic infrastructure refresh, and minimal active support.

Computing Resources

Phase I will use the current SCDF computing infrastructure, which is gradually nearing end-of-life. Around 2030, a refresh is planned to support Phase II, maintaining about 130 kHS06 CPU capacity and 30 PB of disk storage. This smaller hardware footprint aligns with the reduced workload and staffing of Phase II and helps ensure sustainable operations. Funding for this refresh will be requested from the DOE.

Planned Computing Resources:

- 2025–2030: Declining existing capacity – adequate for Phase I
- 2031+: ~130 kHS06 CPU, ~30 PB disk – refresh for Phase II, sustainable with periodic updates

Risks and Mitigation

Key risks include maintaining sustained funding, ensuring active participation from experiments, losing key personnel, capturing knowledge before experts leave, and securing institutional commitments. Mitigation strategies involve early DOE engagement, formal agreements with contributing institutions, phased implementation that allows adjustments, and ongoing monitoring of metadata and workflow progress.

Priorities for 2026

In 2026, the DAP Plan will focus on building the core team, expanding AI-driven knowledge capture across all experiments, and starting coordinated metadata curation. Integrating DAP practices with experiment publication processes will ensure datasets are properly documented and preserved. Metadata standards will be created for cross-experiment discoverability, and prototype containerized workflows will demonstrate reproducible analyses, laying the groundwork for broader deployment in later Phase I stages.