RHIC Data and Analysis Preservation Round Table

06/17/2025

Introduction & some notes from previous meeting

Introduction

- **Unusual meeting time:** thank you for accommodating
- **Follow-up** on topics from previous meetings:
  - Institutional dependencies
  - Effort and resource needs
- **Update:** A draft of the **Data and Analysis Preservation Plan (DAPP)** is now available
- **Next Step:** External **DAPP Review is planned**

Institutional Dependencies

BNL IT Division Services

- Indico event and conference management platform
- Email systems and mailing list infrastructure
- Single Sign-On (SSO) authentication systems
- Video conferencing services
- High-speed internal and external networking infrastructure
- Cybersecurity monitoring and incident response

RHIC-Funded Services (SCDF Managed)

- Computing resources for data processing and analysis
- Enterprise-grade storage systems with hierarchical storage management
- Backup and disaster recovery infrastructure
- Database hosting and management services
- Authorization systems
- Code repository and configuration tools
- Technical support and operational oversight

Computing and Data Science Directorate

- Access to state-of-the-art computing resources for AI applications
- Related technical expertise

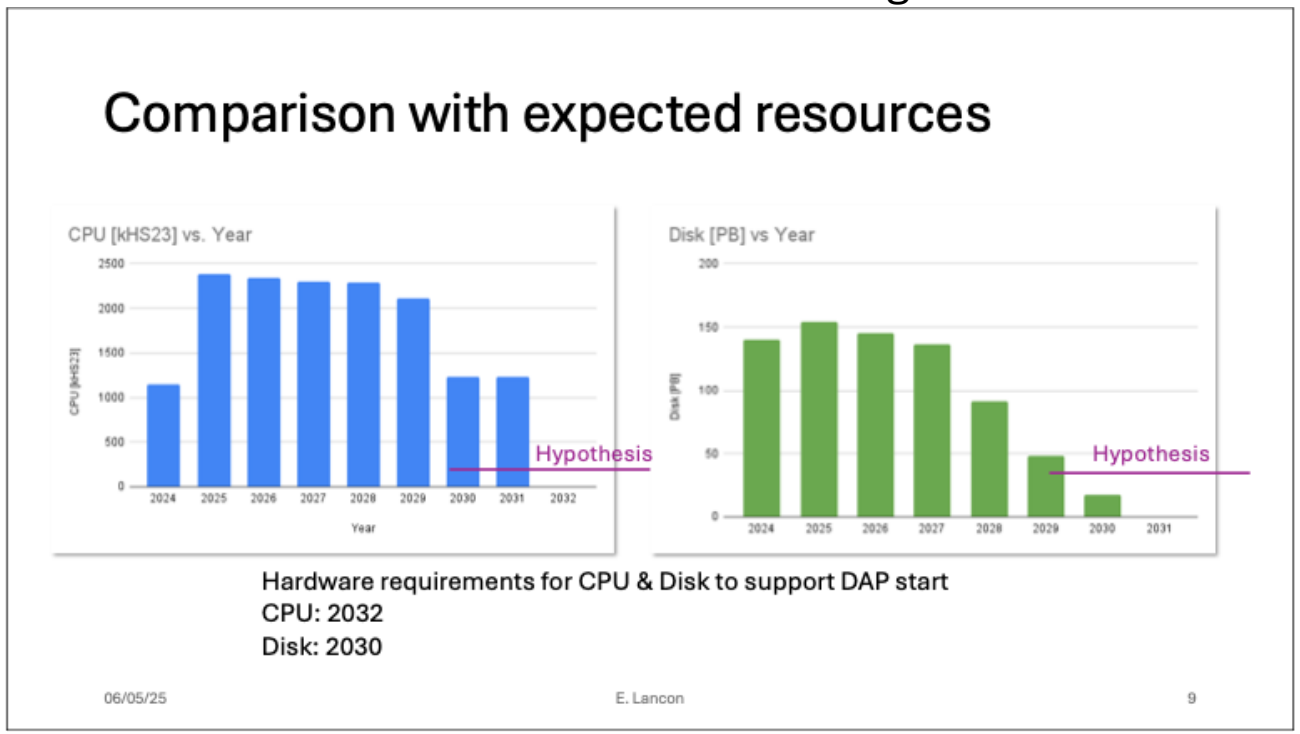
Anything missing?

This dependencies have associated risks

Hypothesis for resources in DAP mode

- Parameters can be adjusted
- **CPU:**
  - The hypothesis is that activity will be dominated by analysis
  - Todays STAR analysis is ~50% of CPU consumption
  - Assume : twice of today's STAR CPU consumption for analysis that is 260 kHS23
- **Hot storage**
  - Rely on tape as backup and carousel for efficient retrieval
  - Assume 30 PB
  - Cannot afford to store the full Analysis Object data sample
  - Rely on Data Carrousel

Slide from 06/05 meeting



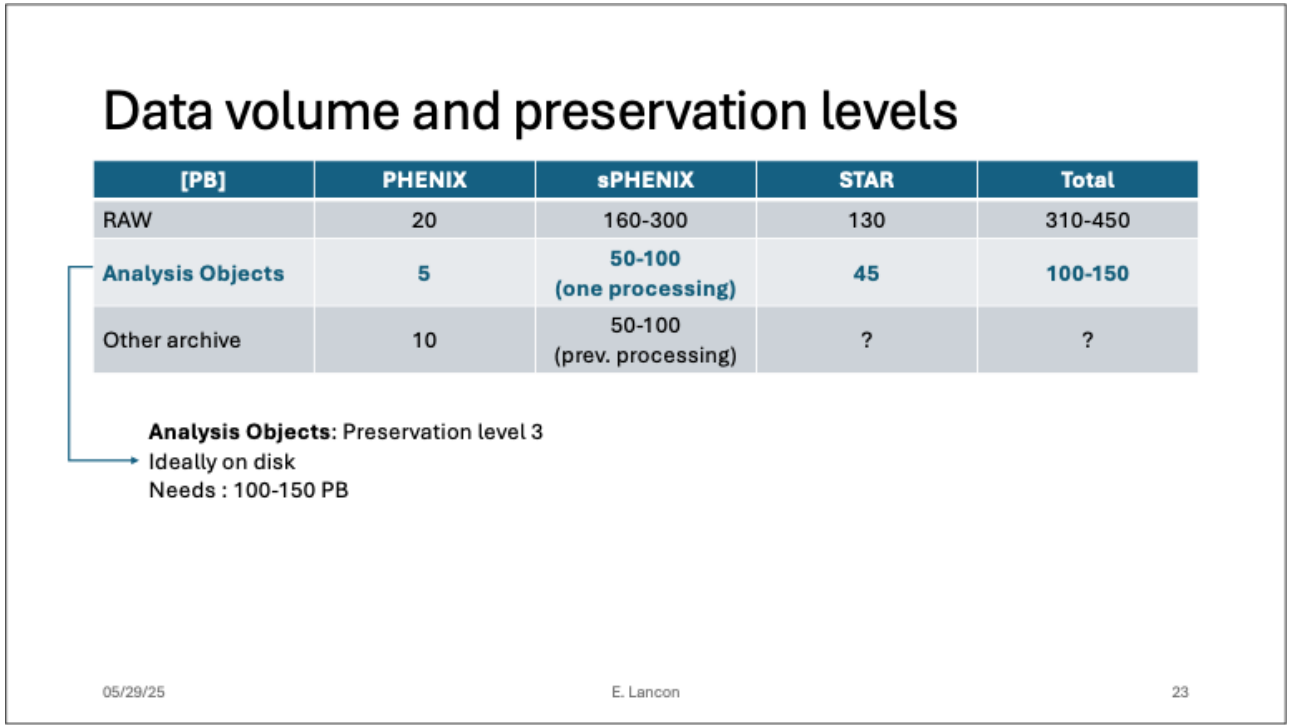
Those are parameters based on hypothesis

Tape systems

Preliminary estimates

- One copy of RAW and
- One copy of Analysis Objects
- RAW: 585 PB (from HPC archive + run 2025 estimate)
  Analysis Objects: 195 PB
- 45 + 15 tape drives

Slide from 05/29 meeting



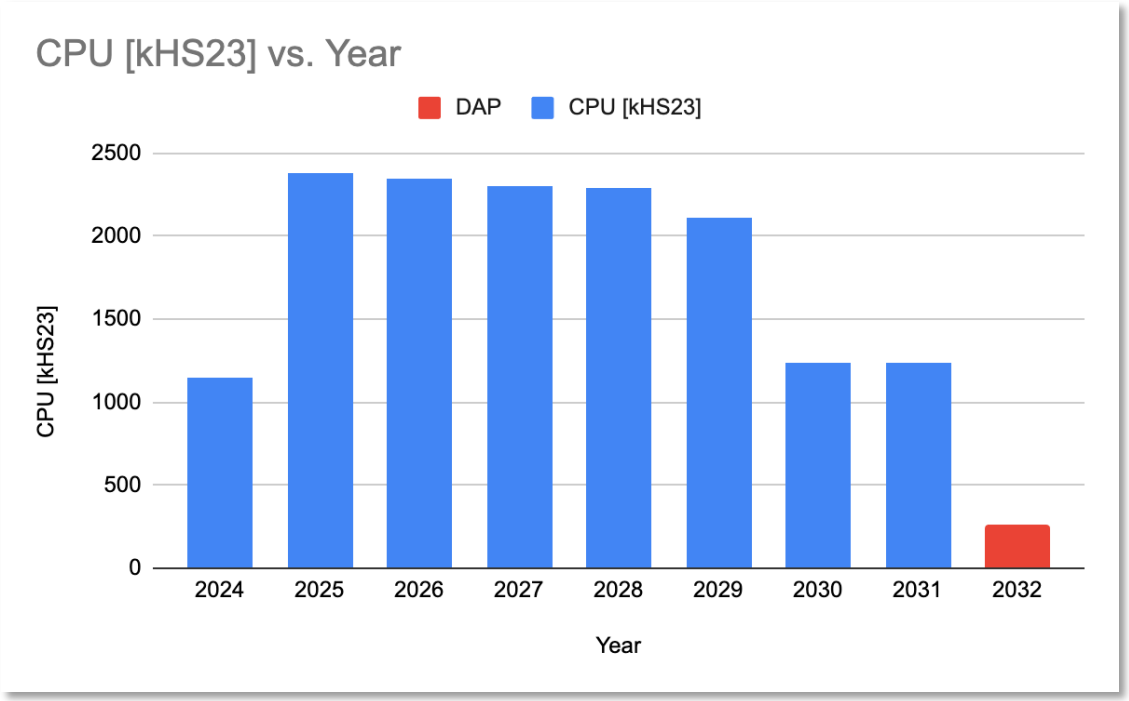
Data volume and preservation levels

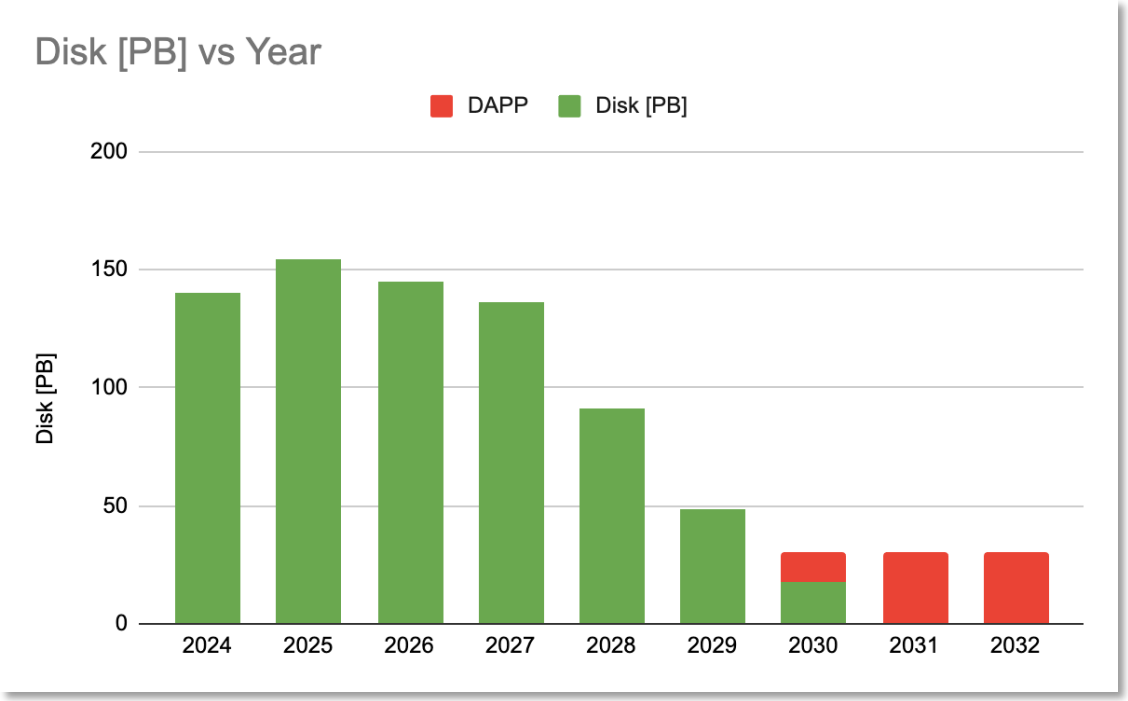
[PB]	PHENIX	sPHENIX	STAR	Total
RAW	20	160-300	130	310-450
Analysis Objects	**5**	**50-100 (one processing)**	**45**	**100-150**
Other archive	10	50-100 (prev. processing)	?	?

Analysis Objects: Preservation level 3
Ideally on disk
Needs : 100-150 PB

05/29/25 E. Lancon 23

Time evolution of resources

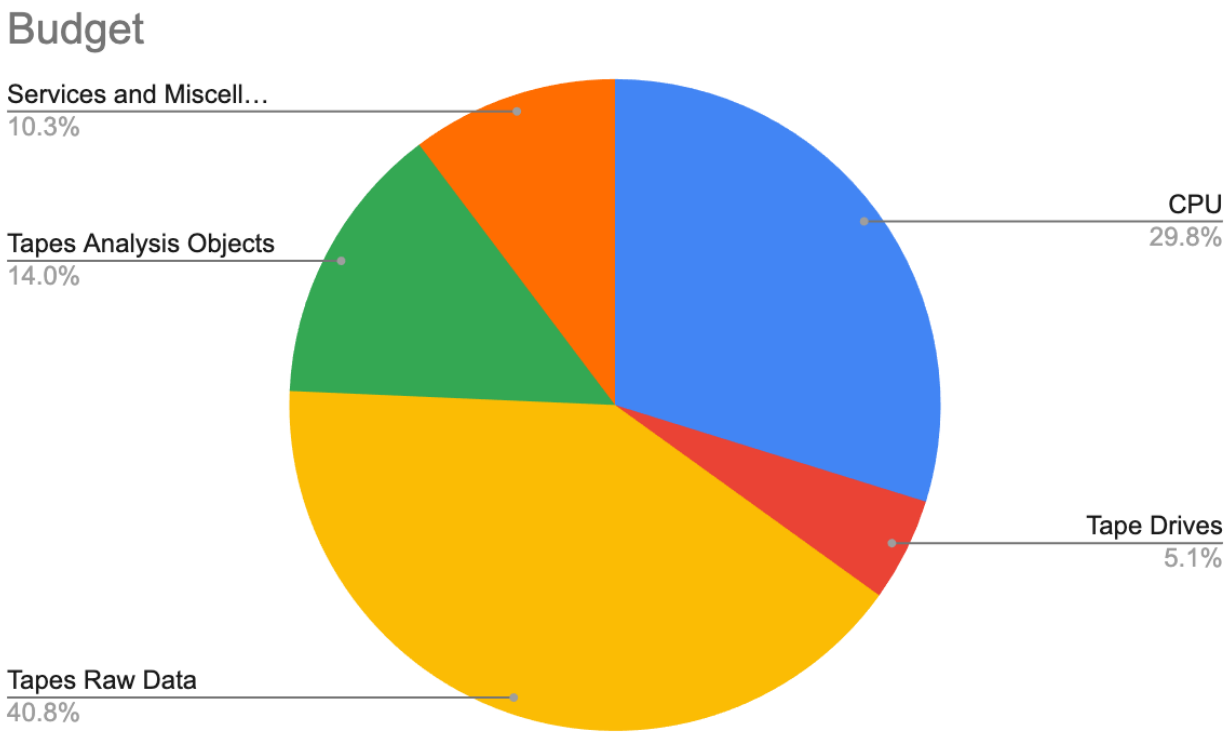




Next refresh:
- 2035 for Disk
- 2039 for CPU

DAPP estimated hardware budget

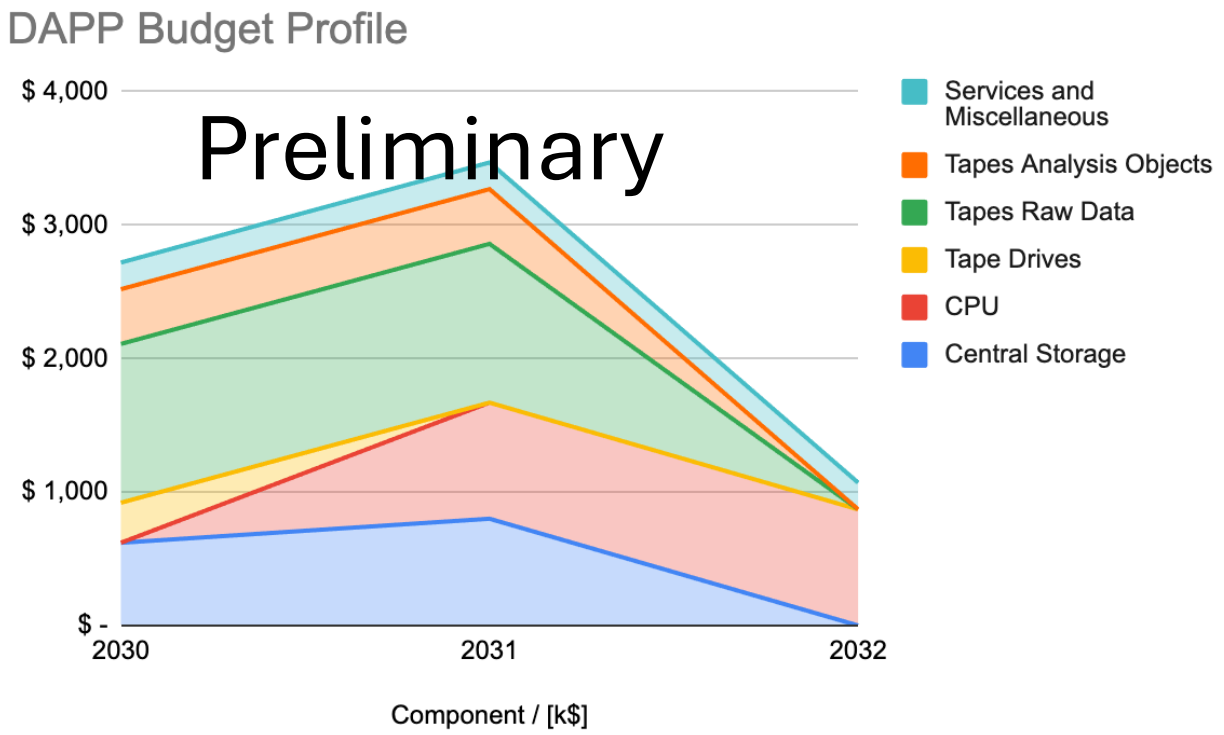
- [2030-2032] Plan
- With this budget basic DAP infrastructure (CPU, Disk, Tape) is setup



Preliminary hardware budget

These are Preliminary numbers

Hypotheses will be refined as Phase II approaches… So these are not final numbers.



RHIC Data Preservation Roles

The Preservation Coordinator plays a special and important role within each experiment

Core Team

- DAP Manager
- Software & Workflow
- Repository Systems
- AI Integration
- Web Development
- Documentation & QA

Experiments

- **Preservation coordinator**
- Scientific Knowledge Preservation
- Code Preservation
- Engagement & Training
- UX & Documentation Design
- Impact Analysis

Computing Center

- Computing Center Liaison
- Technology Watch

From last meeting

Draft of Data and Analysis Preservation Plan about to be released…

It is now available…

Slide from 06/05 meeting

Draft of the DAP Plan (the DAPP)

- **Overview:**
  - A **first draft** of the DAPP has been prepared.
  - It captures ideas and priorities discussed during the RHIC DAPP roundtable series.
  - The document was initially expected to be released by this summer.
- **Next Steps:**
  - The draft will be shared shortly.
  - Everyone is encouraged to review and provide feedback.
  - Dedicated time will be allocated in upcoming meetings for discussion and refinement.


This document is evolving as the preservation plan develops.

RHIC Data and Analysis Preservation Plan

Executive Summary

The Relativistic Heavy Ion Collider (RHIC) Data and Analysis Preservation Plan (DAPP) will safeguard more than two decades of DOE-supported scientific data and research investment as RHIC transitions into its legacy phase. The plan will establish a sustainable, trustworthy digital repository to ensure that RHIC's unique datasets; probing the behavior of nuclear matter under extreme conditions; remain accessible and usable by future researchers, educators, and the broader scientific community.

DAPP follows a phased implementation strategy grounded in established digital preservation standards, including the CoreTrustSeal guidelines, and FAIR principles. Phase I (Years 1–5) focuses on building the necessary infrastructure, capturing datasets and analysis tools, and establishing governance and access frameworks while RHIC experimental teams remain actively engaged. This phase may conclude in a comprehensive data reprocessing to create standardized data products. However, significant uncertainties regarding 2025 run data volumes and processing requirements may extend Phase I beyond the initial timeline.

Phase II (Year 6 onward) transitions to long-term stewardship, requiring fewer resources while maintaining data accessibility as computing infrastructure is scaled down. The preservation strategy prioritizes curated, analysis-ready data and core analysis environments, with limited reconstruction capabilities for select simulated and real data workflows. This approach maximizes scientific value while managing long-term costs, recognizing that large-scale reprocessing of raw data is not sustainable.

Resource planning accounts for current uncertainties in data processing and storage needs, which will be better understood and refined during Phase I implementation. To enhance long-term usability and broaden scientific impact, the plan incorporates AI-assisted tools to support data discovery, contextual interpretation, and navigation of complex analyses. These tools aim to reduce barriers for new users—particularly those without prior RHIC experience—and ensure continued return on DOE's investment in RHIC science.

Strategic Vision and Context

Introduction

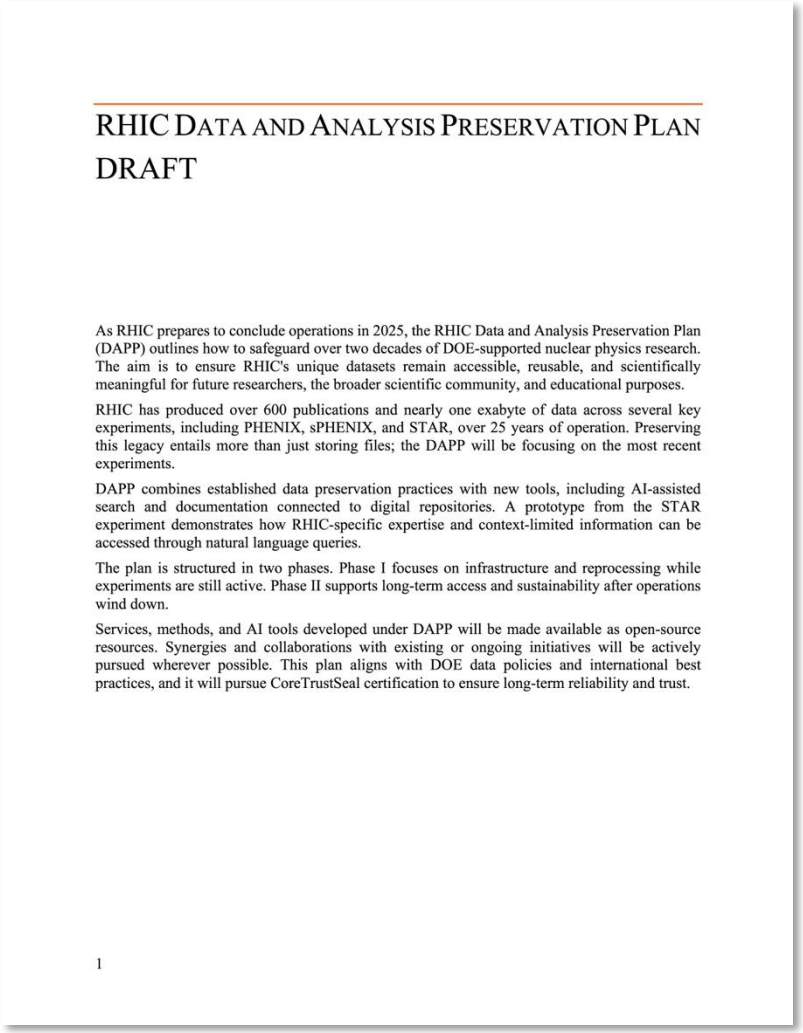
Since beginning operations in 2000, the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory has generated groundbreaking insights into the fundamental nature of matter. The facility's three major experiments; PHENIX, STAR, and sPHENIX; have collected hundreds of petabytes of data by colliding heavy ions at unprecedented energies.

1

DRAFT

06/05/25 E. Lancon 12

Updated draft is available Here



RHIC DATA AND ANALYSIS PRESERVATION PLAN DRAFT

As RHIC prepares to conclude operations in 2025, the RHIC Data and Analysis Preservation Plan (DAPP) outlines how to safeguard over two decades of DOE-supported nuclear physics research. The aim is to ensure RHIC's unique datasets remain accessible, reusable, and scientifically meaningful for future researchers, the broader scientific community, and educational purposes.

RHIC has produced over 600 publications and nearly one exabyte of data across several key experiments, including PHENIX, sPHENIX, and STAR, over 25 years of operation. Preserving this legacy entails more than just storing files; the DAPP will be focusing on the most recent experiments.

DAPP combines established data preservation practices with new tools, including AI-assisted search and documentation connected to digital repositories. A prototype from the STAR experiment demonstrates how RHIC-specific expertise and context-limited information can be accessed through natural language queries.

The plan is structured in two phases. Phase I focuses on infrastructure and reprocessing while experiments are still active. Phase II supports long-term access and sustainability after operations wind down.

Services, methods, and AI tools developed under DAPP will be made available as open-source resources. Synergies and collaborations with existing or ongoing initiatives will be actively pursued wherever possible. This plan aligns with DOE data policies and international best practices, and it will pursue CoreTrustSeal certification to ensure long-term reliability and trust.

1

Table of Content

Feedback !

- Please read the document and provide feedback and comments
- Time is short…
- A review is coming…

Review of DAPP

Chair:

Cristel Diaconu – H1 / DPHEP Chair

Members

- Simone Campana – CERN
- Achim Geiser – ZEUS / CMS
- Kati Lassila-Perini – ICFA Data Life Cycle
- Ulrich Schwickerath – CERN Preservation
- Ralf Seidl – BNL / RIKEN

Date

- **July 1st**
- Half-day (morning session)
- **Over Zoom**

Nuclear and Particle Physics Directorate

Brookhaven National Laboratory

Building XXX
P.O. Box 5000
Upton, NY 11973-5000
Phone 631. 344.8783
abhay@bnl.gov

managed by Brookhaven Science Associates
for the U.S. Department of Energy

www.bnl.gov

Abhay Deshpande
Associate Laboratory Director
Nuclear and Particle Physics

June 10, 2025

Dr. Christinel Diaconu
Directeur de Recherche CNRS
Director of Centre de Physique des Particules de Marseille

Dear Dr. Diaconu:

As the Relativistic Heavy Ion Collider (RHIC) enters its legacy phase after more than two decades of groundbreaking nuclear physics research operation, Brookhaven National Laboratory has started developing a comprehensive Data and Analysis Preservation Plan (DAPP) for the data collected over the years. At this stage of the data preservation project, we are seeking expert input to validate the approach and ensure its alignment with long-term scientific vision of preserving the data, software for at least two decades.

With this letter, I am requesting your committee to conduct a half-day internal review of this plan. The suggested time is July 1, 2025 starting at 9:00AM Eastern US time. A brief written report summarizing the committee's findings and recommendations would be highly appreciated within one week following the review. The DAPP team will deliver a detailed presentation and will be available to address your questions and receive your feedback.

The committee's evaluation should focus on the following questions:

1. Has the DAPP effectively identified and plan to preserve the most valuable scientific assets and legacy from the RHIC experiments?
2. Will the proposed infrastructure enable both verification of published results and new analyses by external researchers?
3. Are proposed data curation practices sufficient to ensure long-term usability and discoverability of RHIC data?
4. Are the proposed FTE allocations and infrastructure requirements realistic for both the initial and sustained implementation phases?
5. Has the plan identified risks and outlined suitable mitigation strategies?

Charge questions

The committee's evaluation should focus on the following questions:

1. Has the DAPP effectively identified and plan to preserve the most valuable scientific assets and legacy from the RHIC experiments?
2. Will the proposed infrastructure enable both verification of published results and new analyses by external researchers?
3. Are proposed data curation practices sufficient to ensure long-term usability and discoverability of RHIC data?
4. Are the proposed FTE allocations and infrastructure requirements realistic for both the initial and sustained implementation phases?
5. Has the plan identified risks and outlined suitable mitigation strategies?

Thank you