RHIC Data and Analysis Preservation Plan

Response to 2024 PAC Recommendation

Eric Lancon elancon@bnl.gov

PAC – October 2025

@BrookhavenLab

PAC 2024 Request & Our Response

*"The PAC requests that BNL Management provide at the next PAC Meeting a detailed report on the **status and plans** for long-term RHIC data analysis and data preservation, including **workforce, computing resources**, and with possible **timelines**."*

Our Response:

- ✓ Comprehensive Data and Analysis Preservation Plan (DAPP) (June 2025)
- ✓ Detailed workforce, computing, and timeline plans developed
- ✓ Independent expert review (July 2025)

June 2025

RHIC DATA AND ANALYSIS PRESERVATION PLAN

As RHIC prepares to conclude operations, 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 more than 600 publications and nearly one exabyte of data over 25 years of operation. Preserving this legacy entails more than just storing files; the DAPP focuses on maintaining access to experiment data, associated knowledge, and analysis workflows in a usable and interpretable form.

The plan integrates established data preservation practices with modern tools, including AI-assisted documentation and search capabilities connected to digital repositories. These tools aim to improve discoverability and long-term usability while being designed to support RHIC-specific contexts and workflows.

The effort is structured in two phases. Phase I emphasizes infrastructure development and data reprocessing while experiments are still active. Phase II supports long-term access, stewardship, and sustainability after operations conclude. Services, methods, and tools developed under DAPP will be released as open-source resources, and synergies and partnerships with ongoing initiatives will be actively pursued. This plan aligns with DOE data policies and international best practices, and it will pursue CoreTrustSeal certification to ensure long-term reliability and trust.

For this plan to succeed, active participation from the RHIC collaborations is essential, along with sustained support for research at participating institutions to maintain the necessary expertise and engagement over time.

Finally, the tools and processes developed through DAPP could serve as valuable models for other large-scale scientific programs, offering transferable solutions for long-term data preservation, access, and reuse.

1

This presentation summarizes progress and readiness for implementation

Independent Expert Review – Strong Endorsement

> *"Given the importance of the scientific assets involved, the robustness of the proposed strategy and the adequacy of the requested resources, the committee strongly recommends this project for support by the funding agencies."*

Committee Recommendation (July 2025)

Review Committee:

- Cristinel Diaconu (Chair, CPPM, CNRS/IN2P3)
- Kati Lassila-Perini (University of Helsinki)
- Achim Geiser (DESY)
- Simone Campana (CERN)
- Ralf Seidl (RIKEN)
- Ulrich Schwickerath (CERN)

Key Findings:

✓ Robust and forward-looking strategy
✓ Realistic resources and phased plan
✓ Innovative AI integration

Endorsed as robust, forward-looking, and aligned with DOE policy.

DAPP Overview

Safeguarding Data, Software, and Knowledge

- Two-phase plan ensures scientific usability long after operations end
- Integration of DOE data policy, FAIR principles, and AI-driven tools
- Open-source approach; adaptable beyond RHIC

Provides a tested U.S. model for NP data preservation

End-to-End Reproduction of Published Results

$ `reproduce Figure 12 from STAR's paper:` Measurement of inclusive charged-particle jet production in Au+Au collisions at $\sqrt{s_{nn}}$ = 200 GeV

User workflow based on AI agents

1. **Portal**: Access the DAP Portal via single sign-on authentication, enter query in natural language
2. **Search**: Locate publication, datasets, and preserved software environment
3. **Dataset**: Open linked, analysis-ready datasets
4. **Container**: Launch container with exact software and configurations
5. **Run**: Execute workflow, reproduce plots and tables
6. **Verify**: Compare results to the original publication

Key advantages

- No setup needed: fully containerized environments
- Full provenance: software, parameters, and data exactly match original work
- Easy to use: AI guidance lowers the barrier for non-experts

Preserve the Ability to Redo Analyses

- Archive all raw data on tape at SCDF
- Preserve reconstruction & simulation software in containers
- Capture complete workflows for reproducibility
- Comprehensive documentation of software, workflows, and metadata captured as part of preservation process
  - Input from experiments is essential to ensure accuracy and completeness
- Focus on analysis-ready datasets (AOD/DST or equivalent)
- Enable targeted small-scale reprocessing if needed

Ensures RHIC data remain scientifically usable long after data taking ends

Two-Phase Plan

Phase I (2026-2030)

Build and Capture

- Infrastructure development
- Knowledge capture while experts available
- Final data reprocessing (possibly)

Staffing: 6.5 FTE
Full computing resources available

Phase II (2031+)

Sustainable Operations

- Long-term data access
- User support & System maintenance
- No bulk reprocessing capacity

Staffing: 2.1 FTE
Reduced computing resources

*"The committee strongly supported the **phased approach as cost-effective and technically sound**."*
— External Review Committee, July 2025

Realistic Expectations and Boundaries

Within Scope

- Long-term access to curated analysis-ready datasets (AOD/DST)
- Verified reproducibility for key published analyses
- Education, outreach, and training use
- FAIR-compliance

Beyond Scope

- Full reconstruction of all raw data
- Preservation without sustained support
- Automatic portability to other institutions
- Recovery of undocumented or lost analyses

The plan focuses on what can be maintained and validated

Major Achievements in 2025

Significant Progress Since Last PAC

DAP Roundtables

Regular DAP roundtables established with PHENIX, sPHENIX, STAR, and external experts

AI ChatBot

ChatBot indexing thousands of technical documents

Document repository

InvenioRDM deployment with DOI assignment and version control

Open Data Portal

CERN Open Data Portal adapted and populated with PHENIX data

Building technical & organizational foundation

Addressing PAC Request – Computing Resources

Phase I (2025-2030)

Implementation

- SCDF declining capacity as hardware ages
- Adequate for Phase I needs
- Global reprocessing campaign possible

Phase II (2031+)

Reduced Capacity

- ~130 kHS06 CPU, ~30 PB disk storage
- Sustainable with periodic updates
- Right-sized for reduced workload

Critical: ~2030 Infrastructure Refresh

- Replace end-of-life hardware before Phase II
- Estimated Budget: **$7.6M** (2030-2032)
- DOE funding request to be submitted

Addressing PAC Request – Workforce Plan

Phase I (2026-2030)

6.5 FTE

Phase II (2031+)

2.1 FTE

*"The committee approved the **DAP core team workforce plan** ..."*
— External Review Committee, July 2025

Detailed Staffing Breakdown

Role	Phase I (2026-2030)	Phase II (2031+)
Coordination & management	1.0	0.5
Software/Workflow	1.0	0.3
Portals & AI	3.25	0.8
Documentation	0.25	0.1
Core Team Total	**5.5**	**1.7**
Experiment Coordinators	1.0	0.3
Total Staffing	**6.5**	**2.1**

Additional Support: Active participation from each experiment (~1+ FTE per experiment during Phase I).
Requires **dedicated resources** to ensure effective engagement from the collaborations.

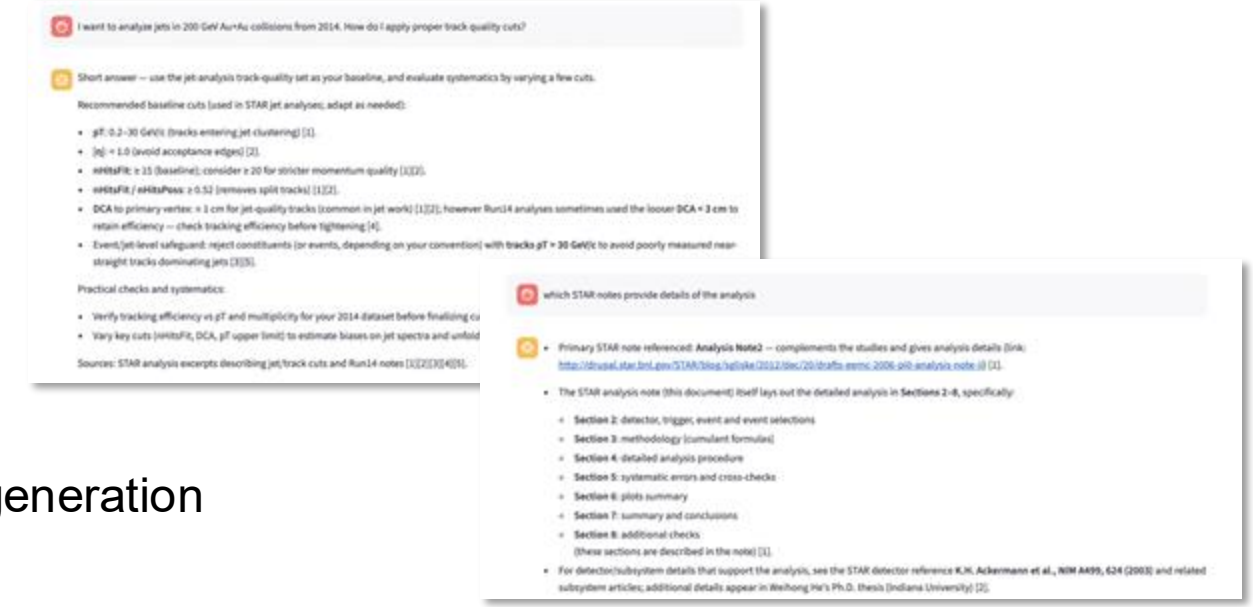
AI-Driven Access to Preserved Knowledge

ChatBot Design:

- LLM core enhanced with **Retrieval-Augmented Generation (RAG)** for precise document access
- Integrates with **Model Context Protocol (MCP)** to manage multiple knowledge sources
- Public LLMs have *no access* to internal notes or repositories

Indexed Sources (Thousands of entries)

✓ Internal notes and documents
✓ Web sites, Indico pages
✓ pdf, doc, pptx, HTML, EPS
✓ …



✓ **STAR prototype deployed and operational**
Next: expand to software documentation and code generation

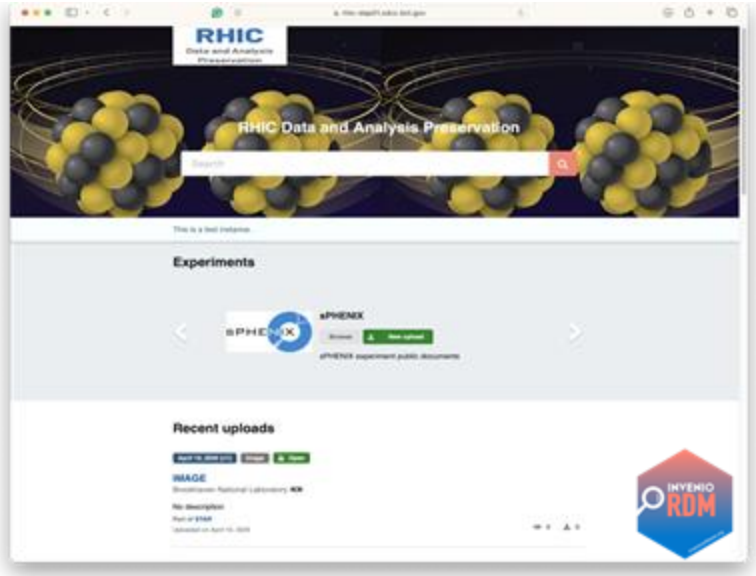
"The committee strongly supported the phased approach as cost-effective and technically sound, ***highlighting the use of AI-driven knowledge navigation****."*
— External Review Committee, July 2025

Document Repository & Data Portal

Document Repository based on Invenio

✓ Multi-experiment hosted at BNL
✓ Version control & provenance tracking
✓ OSTI DOI assignment
✓ Federated authentication
✓ Public & restricted document management

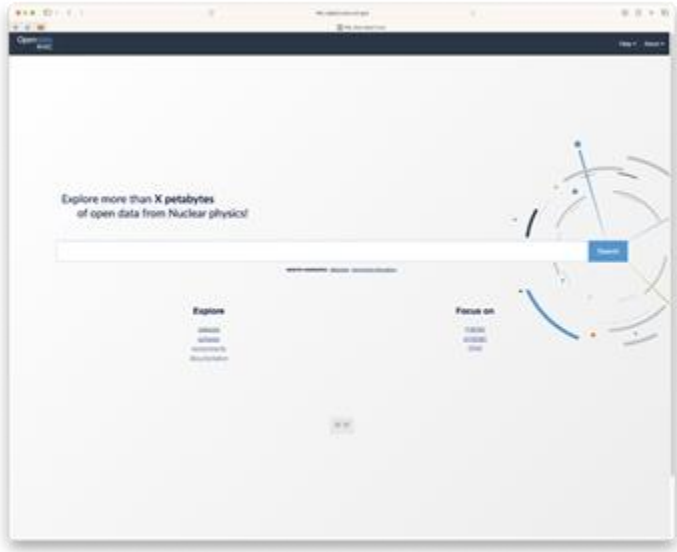
Status: Operational



Open Data Portal

✓ CERN Open Data Portal adapted for BNL
✓ Initially populated with PHENIX datasets
✓ Metadata-based search capabilities
✓ Access restrictions enforced
✓ Search by run, energy, system

Next: STAR and sPHENIX datasets when open data available



Standards & Alignment

- Aligned with DOE data management & security requirements
- FAIR principles: *Findable, Accessible, Interoperable, Reusable*
- Pursuing CoreTrustSeal certification (repository standard)
- International engaged and aligned via DPHEP and through invitations to ICFA Data Lifecycle Panel fostering outreach and collaboration with other communities

FAIR Principles

- **F**indable: Rich metadata, DOIs
- **A**ccessible: Portal, authentication
- **I**nteroperable: Standard formats, APIs
- **R**eusable: Complete documentation

Managing Risks

DAPP: Comprehensive Risk Analysis
16 specific risks identified with mitigation strategies
Active registry updated quarterly since review

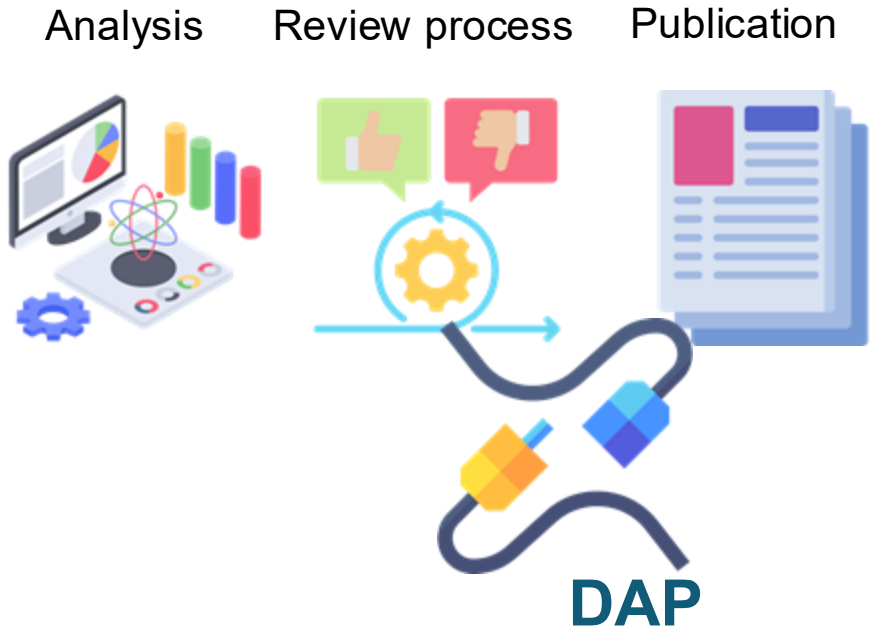
Key Risks & Mitigations

Risk	Mitigation
Funding continuity	Phased approach, early progress
Expertise loss	Knowledge capture + AI ChatBot
Hardware obsolescence	2030 infrastructure refresh
Institutional dependencies	Coordination with BNL SCDF & ITD

2025 achievements significantly reduce implementation risks and demonstrate technical feasibility

2026 Foundation-Building Year

- Establish the core team
- Documentation effort across experiments while teams remain active
  - Substantial effort for documentation and knowledge preservation is needed
- Expand AI-driven knowledge capture across experiments
- Prototype reproducible workflows
- Start coordinated metadata curation
- Integrate preservation steps into publication pipelines



2026 marks the transition from planning to execution.

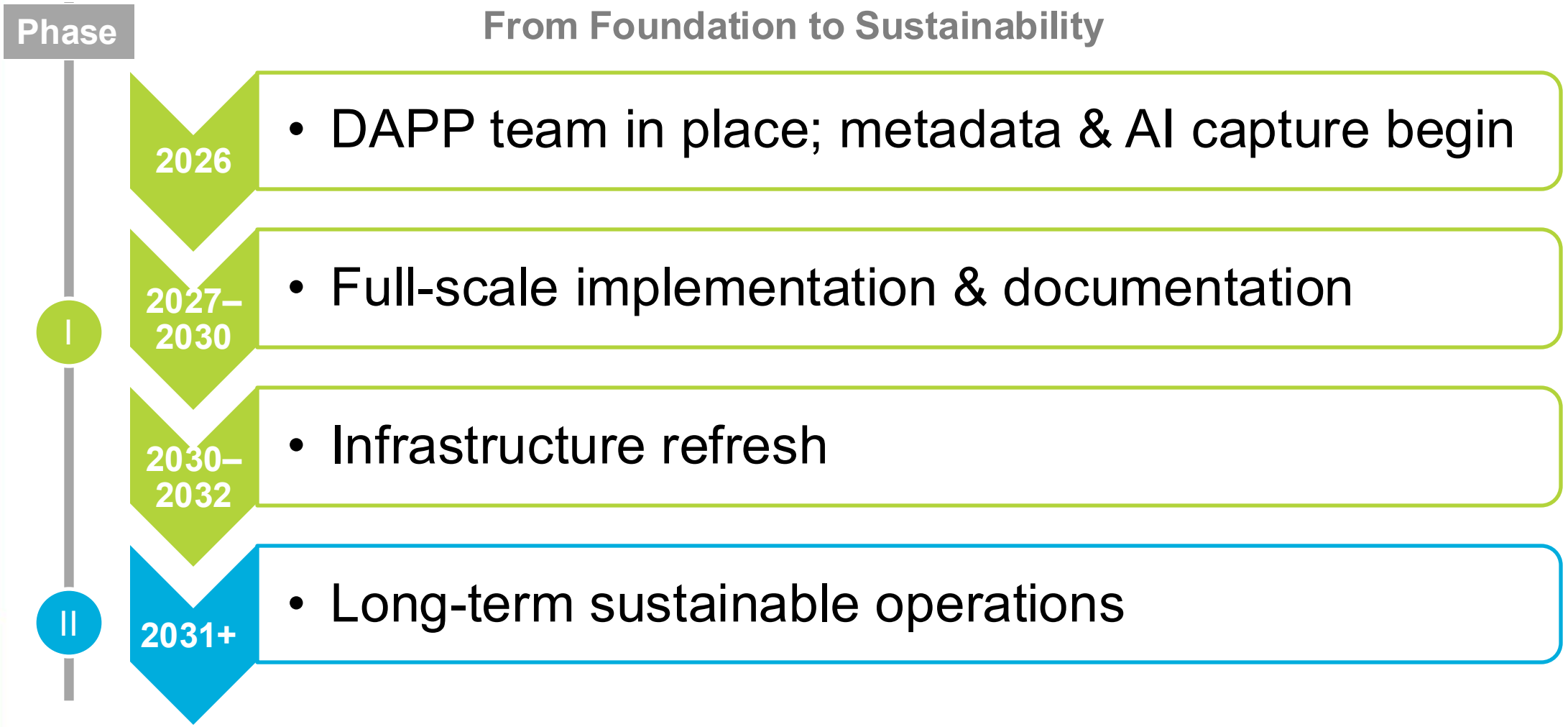
Brookhaven National Laboratory

*"**Metadata curation must begin while expert knowledge is still available** — this is a key challenge emphasized by the review committee."*
— External Review Committee, July 2025

Dependencies for Long-Term Success

Funding Agency	Host Lab	Experiments	Universities & Partners
• Sustained support through both phases • Enable hardware refresh	• Provide computing infrastructure & Institutional services	• Capture expert knowledge & documentation • Validate preserved workflows	• Contribute expertise, training & shared development

Addressing PAC Request – Implementation Timeline



Summary: Ready to Proceed

Protecting RHIC's Legacy for the Future

- ✓ Complete response to PAC 2024 request
- ✓ Endorsed by independent review

Critical Next Steps

Secure DOE funding for both phases

Maintain experiment engagement

Outcome:

DAPP safeguards decades of DOE investment, **extends RHIC's scientific productivity 10+ years**, and provides a tested model for future NP data stewardship.

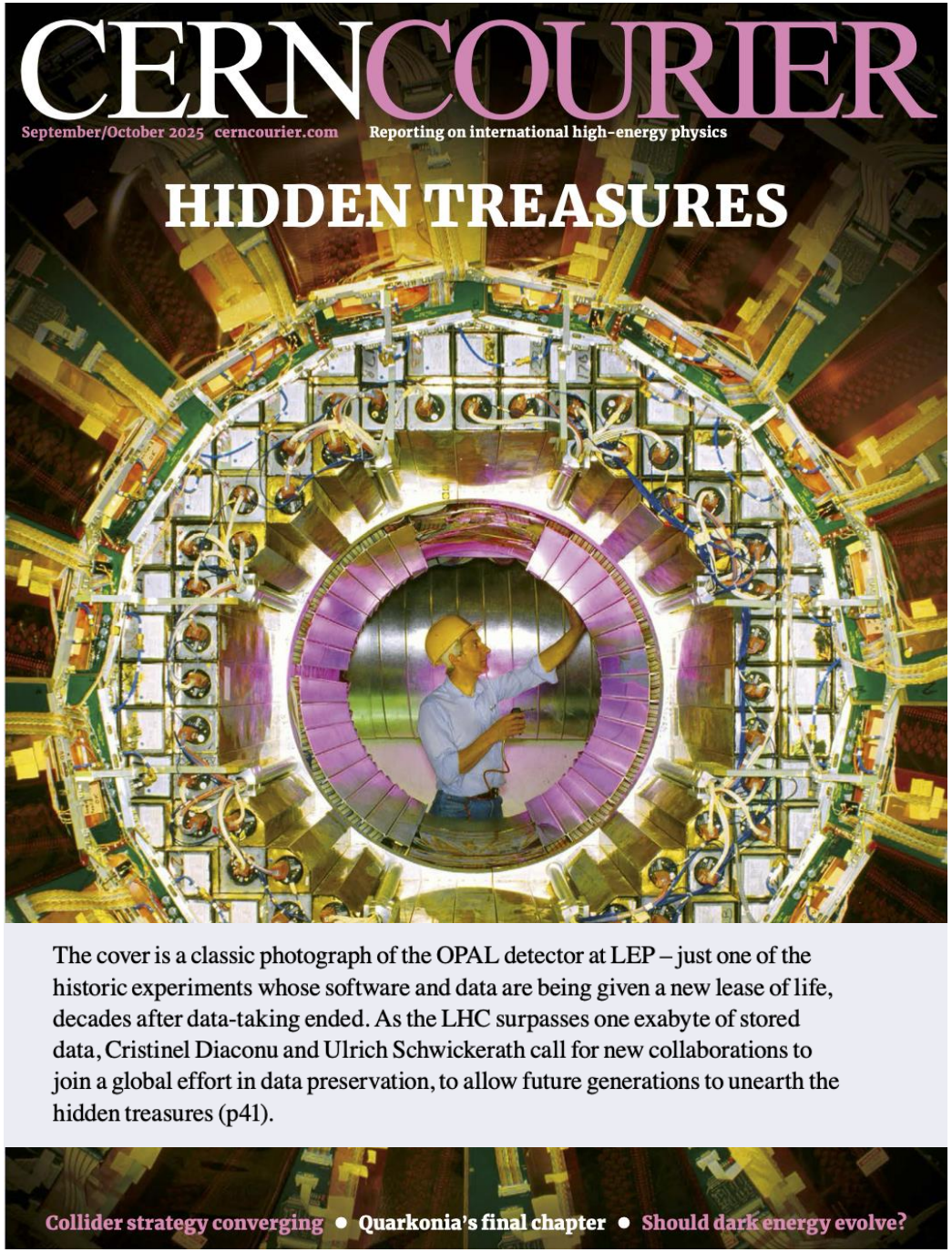
Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

Thank you!

@BrookhavenLab

Thank you!


CERN COURIER
September/October 2025 cerncourier.com
Reporting on international high-energy physics
HIDDEN TREASURES
The cover is a classic photograph of the OPAL detector at LEP – just one of the historic experiments whose software and data are being given a new lease of life, decades after data-taking ended. As the LHC surpasses one exabyte of stored data, Cristinel Diaconu and Ulrich Schwickerath call for new collaborations to join a global effort in data preservation, to allow future generations to unearth the hidden treasures (p41).
Collider strategy converging • Quarkonia's final chapter • Should dark energy evolve?

Backup: Hardware Budget Detail

Total: $7.6M (2030-2032)
For infrastructure refresh

Storage Infrastructure

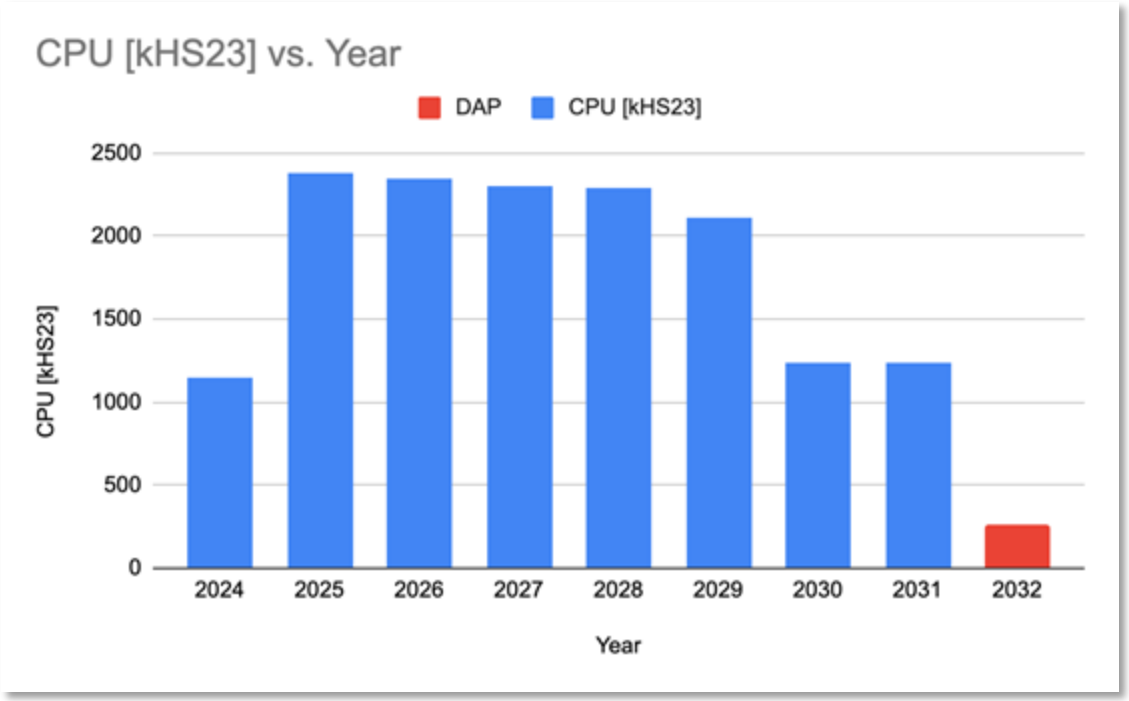
- **Hot Storage:**
  - Current: ~150 PB
  - Phase II: ~30 PB
- **Cold Storage:**
  - LTO-11 migration 2030-2032
  - 60 tape drives, 6,300 cartridges
  - 400-600 PB capacity

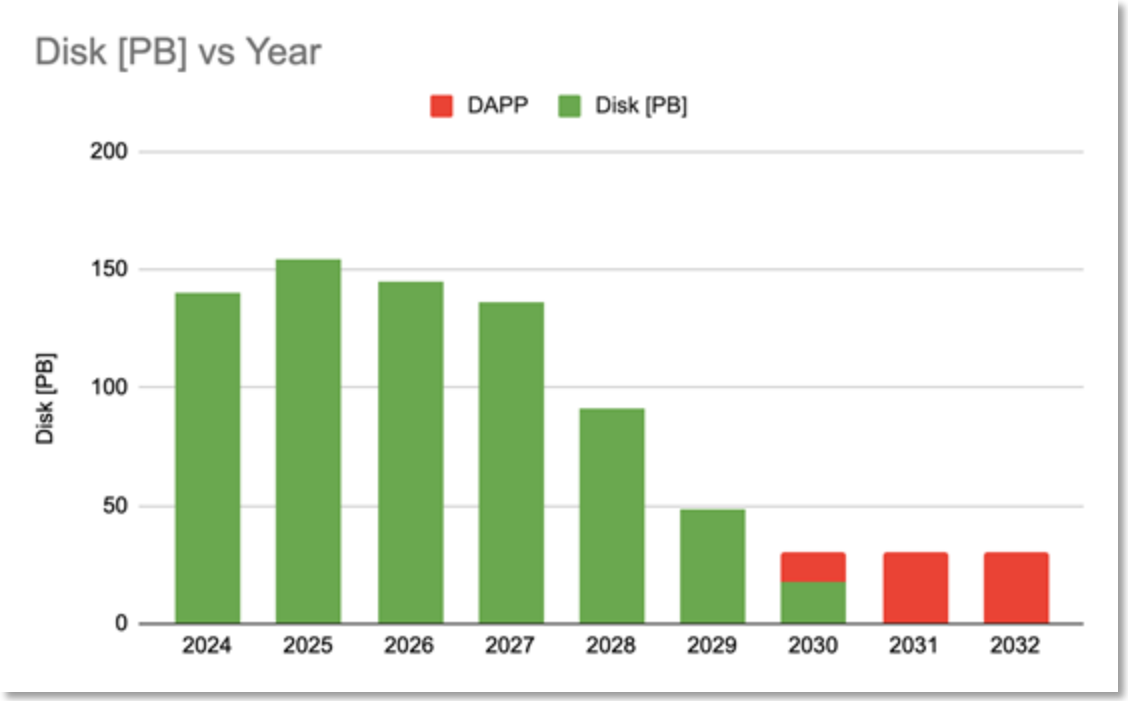
CPU Resources

- **Current Capacity:**
  - 150,000 job slots
  - 7-year hardware lifetime
- **Phase II Requirements:**
  - ~130 kHS06 – twice today's STAR analysis

Component	2030	2031	2032
Central Storage	$620k	$800k	-
CPU Clusters	-	$870k	$870k
Tape Drives	$300k	-	-
Tapes (Raw)	$1,190k	$1,190k	-
Tapes (Analysis)	$410k	$410k	-
Services/Misc	$300k	$300k	$300k
Annual Total	**$2,820k**	**$3,570k**	**$1,170k**

Backup: Time evolution of resources





Next refresh:
- 2035 for Disk
- 2039 for CPU

Backup: Governance Structure

Steering Committee (Oversight)
- Includes RHIC experiments, NPP leadership, external experts
- Meets twice/year (active phase), once/year (maintenance phase)

Technical Working Group (Infrastructure and standards guidance)
- Members from BNL ITD, SCDF, DAPP technical leads
- Advises on infrastructure, preservation standards, best practices

Implementation Team: (Operations)
- Runs roundtables for community engagement and feedback
- Operations and tactical decision-making within strategic framework
- Reports progress to Steering Committee

Transparent, informed, and sustainable project oversight from planning through long-term preservation