RHIC Data and Analysis Preservation Plan (DAPP)

Committee Review Presentation

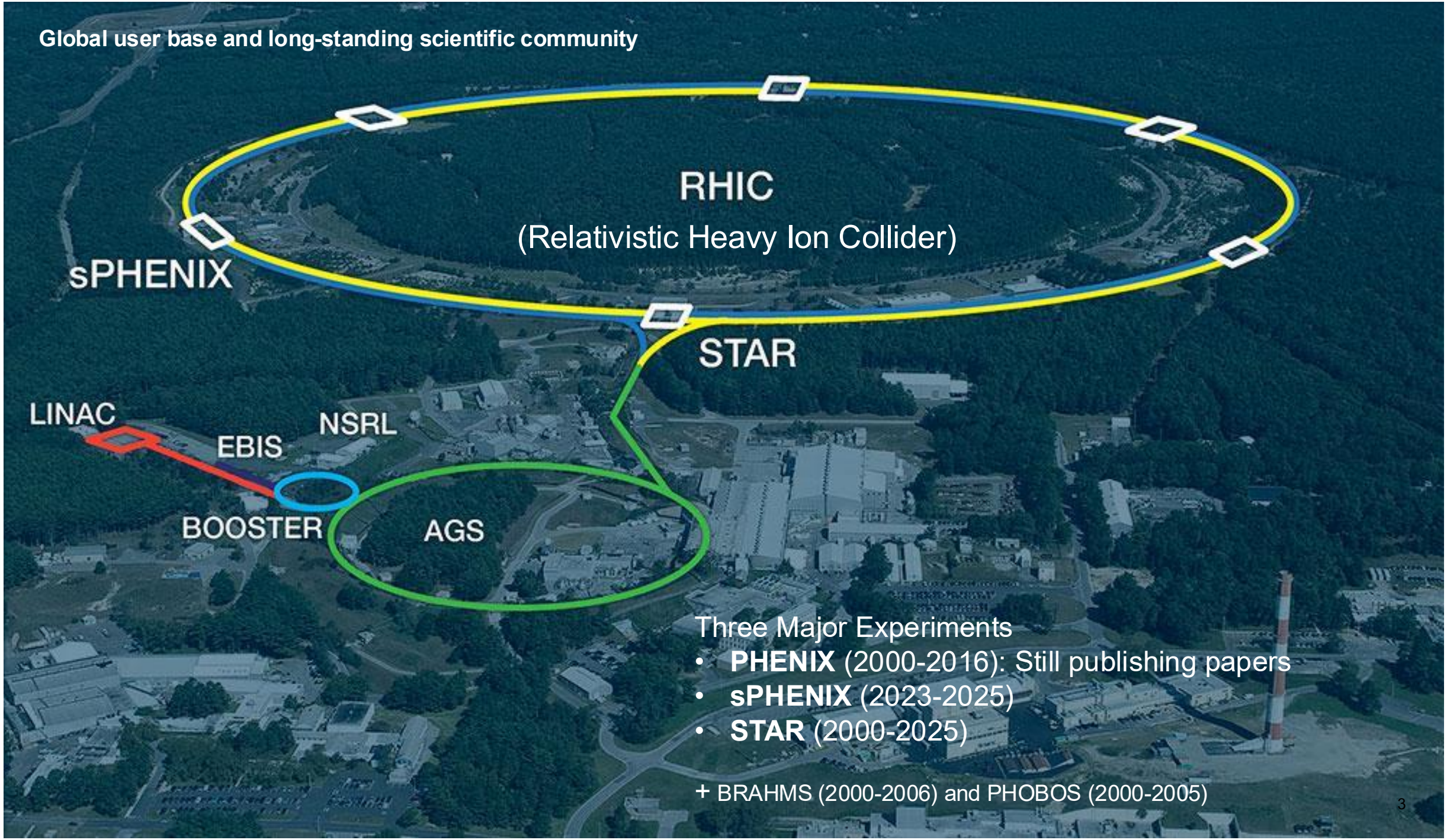
Eric Lancon, Brookhaven National Laboratory

Brookhaven National Laboratory

Agenda

- **Context & Challenges** – RHIC timeline, data volumes, urgency
- **DAPP Strategy** – Preservation approach, and focus areas
- **Technical Platform** – Portal design and infrastructure
- **Implementation Approach** – Timeline and collaboration framework
- **Resource Requirements** – Personnel, hardware, and computing infrastructure
- **Risk Assessment** – Dependencies and mitigation strategies
- **Standards & Compliance** – DOE Order, FAIR principles, regulatory alignment
- **Value & Recommendations** – Scientific impact, success metrics, path forward

Global user base and long-standing scientific community



Three Major Experiments

- **PHENIX** (2000-2016): Still publishing papers
- **sPHENIX** (2023-2025)
- **STAR** (2000-2025)

+ BRAHMS (2000-2006) and PHOBOS (2000-2005)

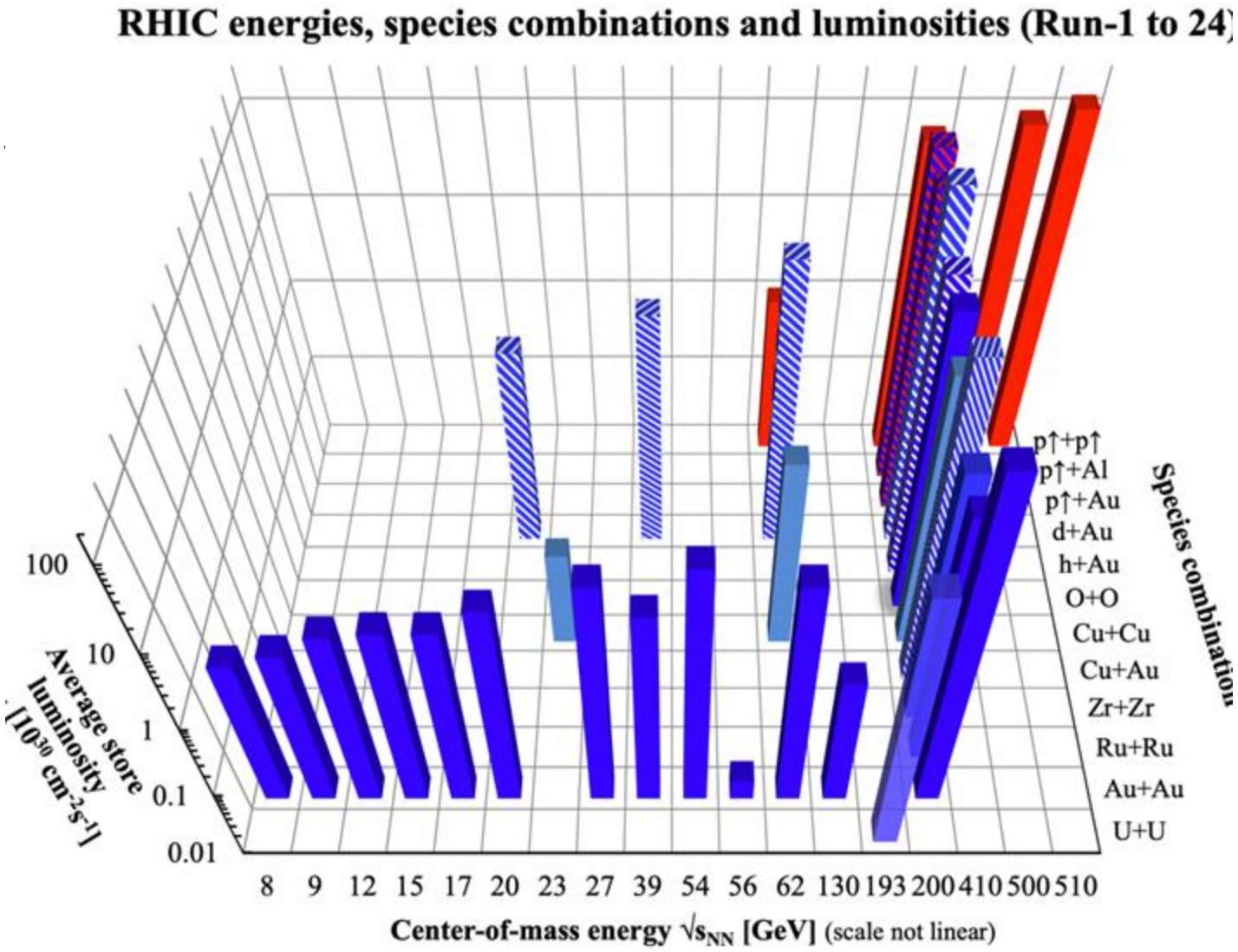
25 years of operations, **Billions invested**, yielding **600+ publications** and **37K+ citations**

25 years of operation

RHIC a versatile facility that can collide:

- **Heavy ions** (gold, copper, uranium, etc.)
- **Polarized protons** to study the internal structure of protons
- **Various combinations** (proton-proton, deuteron-gold, copper-copper, uranium-uranium, aluminum-gold, etc.)
- Also operated in **fixed-target** mode extending the energy range down to 3 GeV

PHENIX stopped taking data, while the other two experiments continue actively taking data, the DAPP needs to be adaptable from the start.



Brookhaven National Laboratory

These data cannot be reproduced elsewhere and represent billions in scientific investment

Context & Key Challenges

RHIC Timeline & DAPP Mission:

- Final operational phase: RHIC shuts down after Run 25
- Expertise will evaporate as collaborations will wind down
- DAPP Mission: Preserve scientific legacy (data, software, expert knowledge)
- Enable future analyses and comparison to theories

Key Challenges:

- Each experiment has different systems and approaches
- The current infrastructure will not be replaced after the end of life of the equipment
- Must complete transition to preservation while expertise is still available
- Large data volume with complex processing requirements
- Au-Au run 2025 will double the collected data volume
- EIC is getting momentum

Urgency: computing, expertise

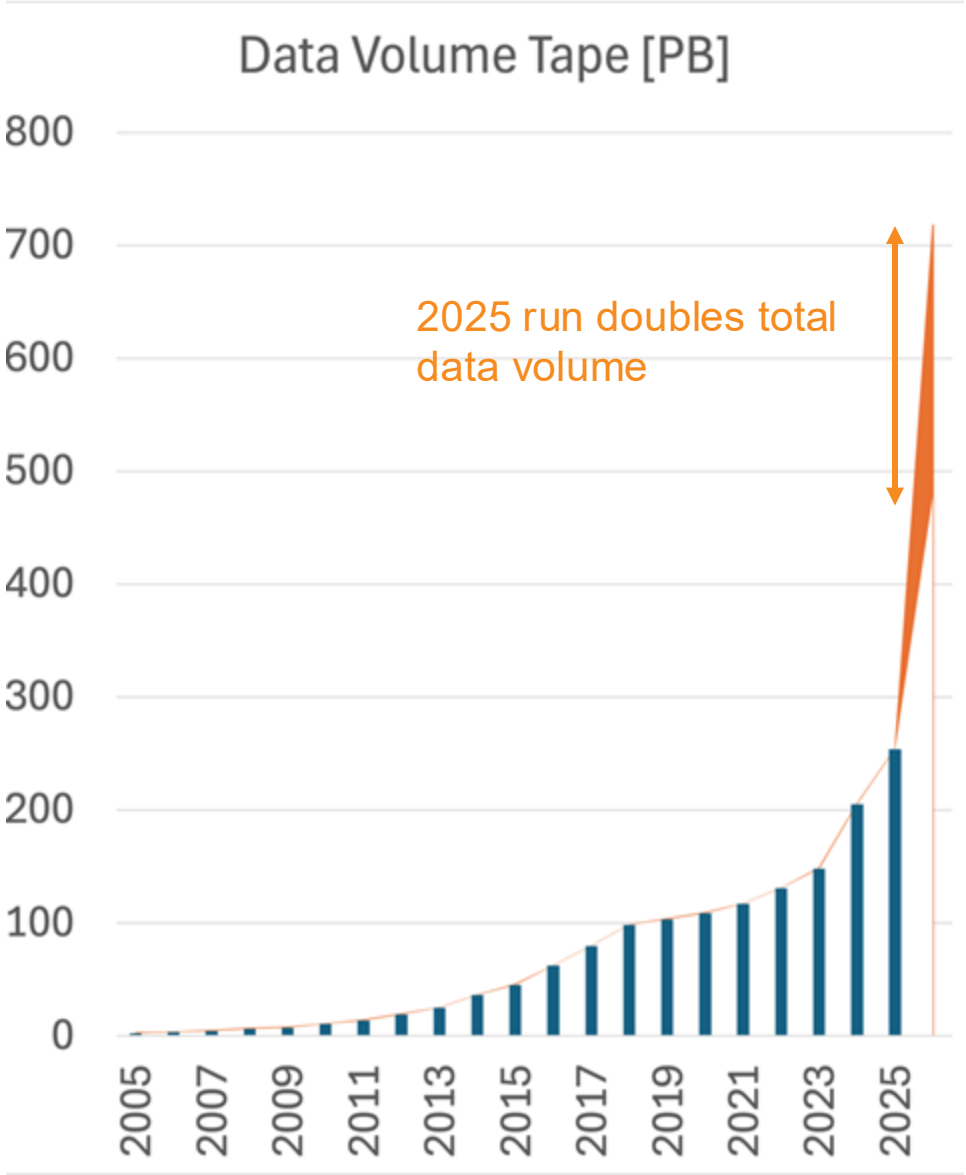
Brookhaven National Laboratory

Lessons from DAPP will directly benefit the upcoming Electron-Ion Collider (EIC), providing a tested model for data and analysis

The Scale Challenge

- Large data volumes
  - Close to 1 Exabyte of total volume (MC is a tiny fraction)
  - Hundreds of Petabytes event for Analysis Objects

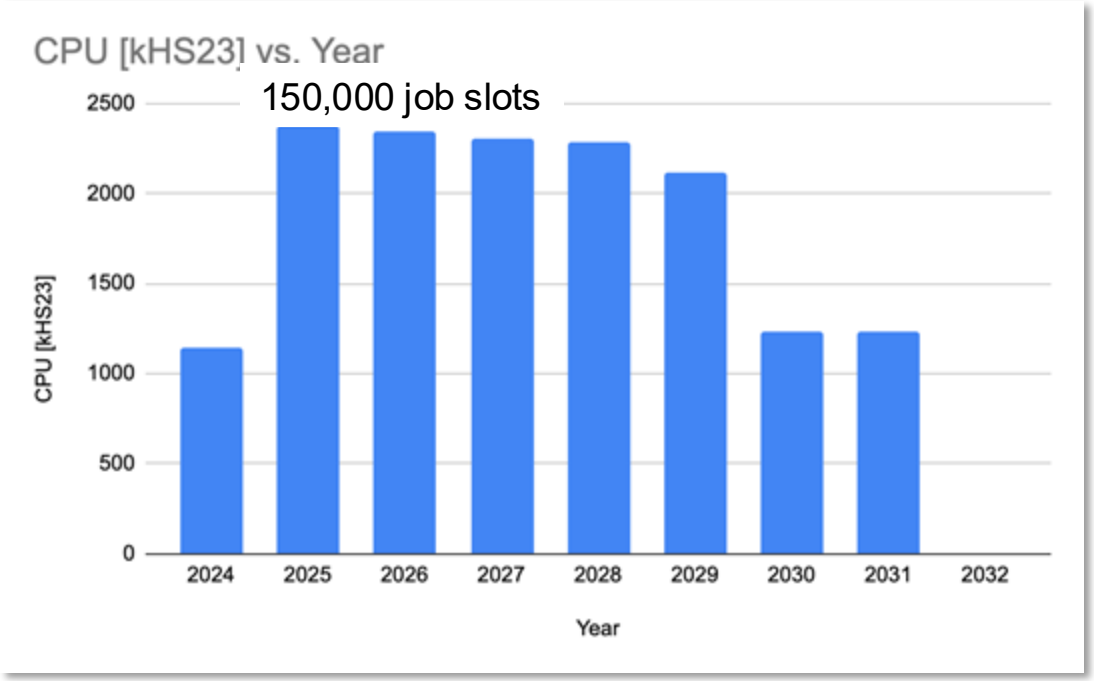
Data Type [PB]	PHENIX	sPHENIX	STAR	**Total**
RAW data	20	160-300	130	**310-450**
Analysis Objects (one processing)	5	50-100	45	**100-150**

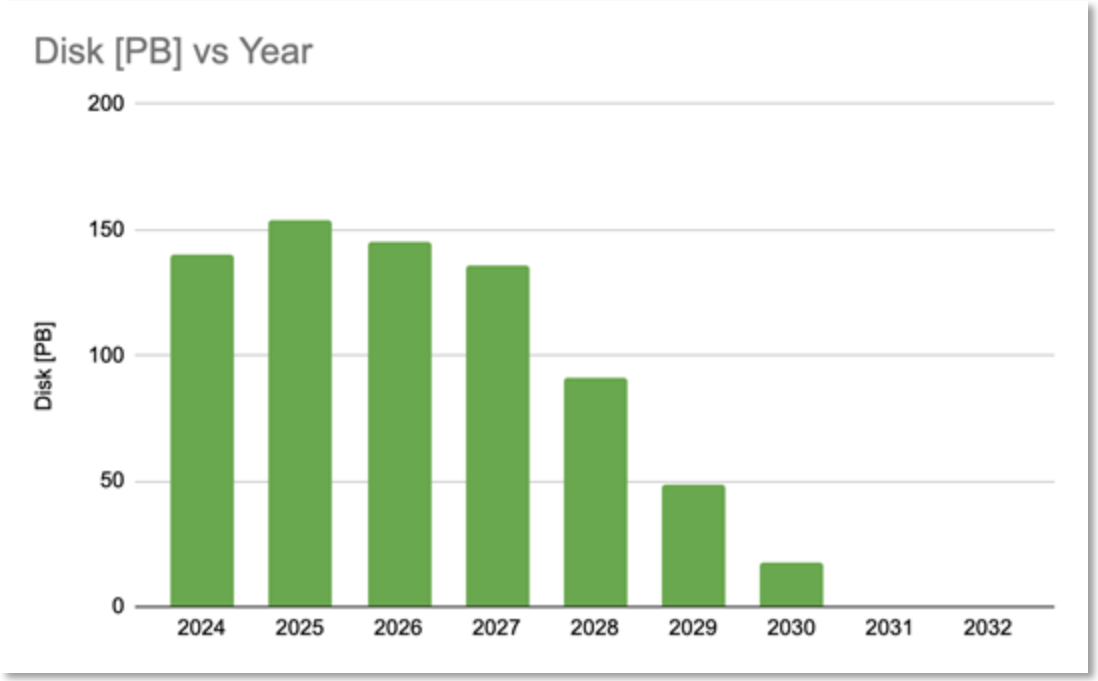


Brookhaven National Laboratory

Evolution of available computing resources

No plan to replace computing infrastructure when end of life reached





Lifetime of
CPU: 7 years
Disk: 5 years

Brookhaven National Laboratory

The Knowledge Transfer Challenge

Experimental Expertise

- Complex experimental knowledge resides within collaborations
- Current scientists will move to other projects (or retire)
- Analysis techniques and software expertise reside within collaborating institutions
- Critical window for knowledge capture closing rapidly

Institutional Engagement

- Institutions hold much of the practical knowledge
- Collaboration members must be actively involved
- Cannot preserve knowledge without those who created it
- Institutional backing enables sustained contributions

Our strategic plan for preservation

Resource-Driven DAPP Approach

Pragmatic Focus

- Emphasize analysis-ready data over reconstructing full raw data.
- Ensure scope is realistic, aligned with available funding and expertise.
- Leverage existing BNL infrastructure
- Use community-developed tools and external partnerships.

Two-Phase Plan

- Phase I: Active development while collaborations remain engaged
- Phase II: Lean operations with reduced but stable funding

Guiding Principles

Multi-User Design Requirements

- Multi-User Design: current collaborators, external researchers, educators, and students
- Ensure intuitive access, even for those new to RHIC

Sustainable Operations

- Reduce long-term maintenance by adopting appropriate standards and technologies
- Allow for graceful degradation: Core functions preserved even under resource constraints
- Prioritize cost-effective, maintainable solutions

Standards & Compliance

- Follow **FAIR** principles: **F**indable, **A**ccessible, **I**nteroperable, **R**eusable
- Implement the **OAIS** model: **O**pen **A**rchival **I**nformation **S**ystem, a reference framework for trusted digital repositories
- Comply with DOE requirements for scientific data management and open science

Scope and Priorities

What's In Scope (Achievable Goals):

- Preserve analysis-ready datasets and associated workflows
- Capture expert knowledge using AI-assisted tools, curated data and documentation
- Maintain environments to allow verification and new analyses based on processed data and simulations.
- Retain the capability for limited access to raw data
- Preservation of simulations and tools (GEANT3) required for reproducibility

What's Out of Scope (Beyond Available Resources):

- Timely reprocessing of large-scale datasets
- Development of new reconstruction algorithms without the original experts.
- Long-term support from experts (many are retiring or moving on)

Preservation Focus

Emphasis: Analysis-Ready Data (~120 PB)

- Processed physics objects (ROOT files)
- Containerized workflows reproducing published results.
- Core documentation: technical notes, detector descriptions, static websites
- Frozen versions of simulation and reconstruction software – preservation is essential to support re-analysis

Commitment: Complete Archive (~700 PB)

- All raw data from RHIC experiments will be preserved on tape
- While full reprocessing is out of scope, preserving the data ensures future capability if priorities change.
- Due to cost and infrastructure requirements, raw data will have less emphasis in terms of resources than analysis-ready products

Two Phases

Phase I (Current – 2030+): Active Implementation – 6.5 FTE

Preparation & Infrastructure Development

Building the systems, tools, and knowledge base needed for long-term preservation **while expert resources are available.**

Foundation Building (Years 1-2), Full-Scale Implementation (Years 3-5)

Phase II (2030+ - No ending date): Sustainable Operations – 2.1 FTE

Sustainable Long-term Operation

Operating with **reduced resources** to maintain accessibility and scientific usability of preserved data

Technical Implementation

RHIC Data Preservation Portal

Single unified access point:

- Intuitive search interface
- Unified access to datasets, documentation, and analysis tools
- Role-based access: public, collaboration, and restricted content
- Built to serve both RHIC experts and new users in research and education
- No software installation required, all tools accessed via containers.

Technical Platform & Infrastructure

Unified Access for different users

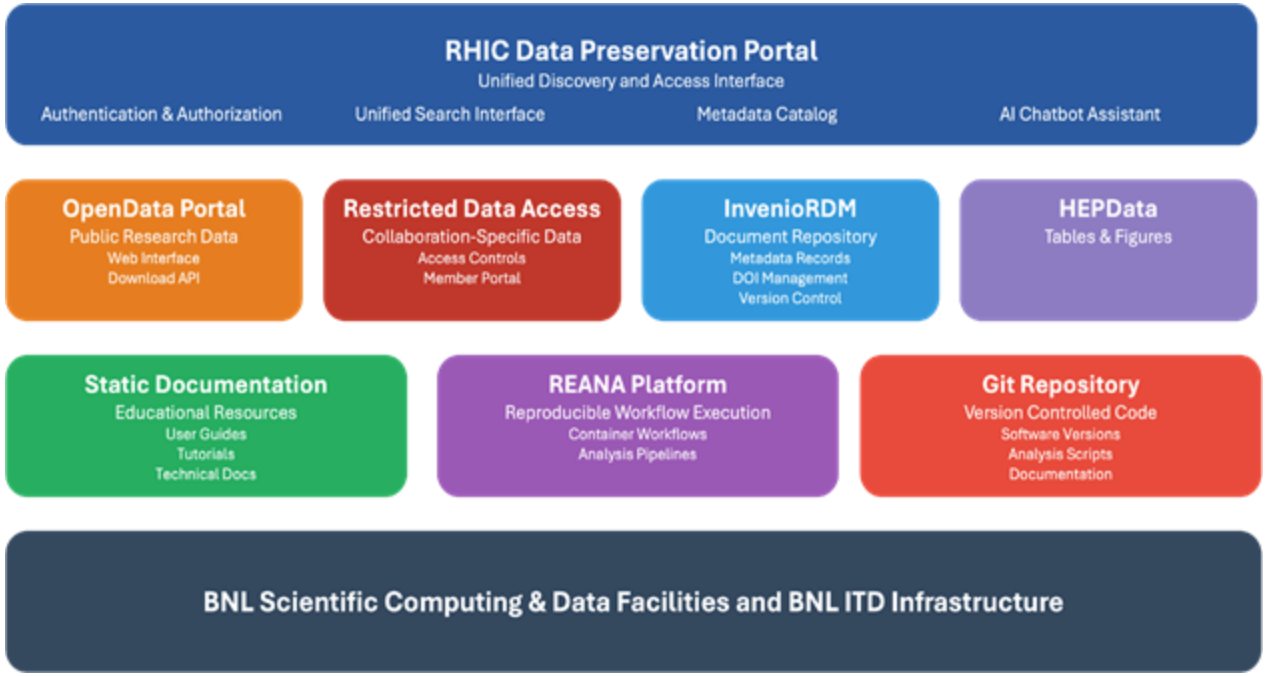
- **Expert Researchers**: Direct access to preserved workflows and technical datasets.
- **New Users**: Guided discovery and reproducible workflows via containerized analysis.
- **Educators & Students**: Interactive examples, simplified datasets, classroom-ready tools

Technical Architecture

- Deployed on OpenShift/Kubernetes for scalability and resilience
- Hosted on BNL's Scientific Data & Computing Facility (SDCF) with ITD support
- Containerized services enable technology evolution

Software Foundation

- Built on the ChatSTAR prototype, extended to support PHENIX, and sPHENIX
- Search built on Retrieval-Augmented Generation (RAG) with Model Context Protocol (MCP)
- Indexes structured content: websites, documents, metadata, software



Data Types, Preservation, Access Paths

Data Type (Input)	Preservation Methods	Access & Reuse
		
RAW Data	Tape storage (1 copy)	Staged retrieval on request
Analysis Objects (ROOT)	Tape + partial disk	REANA containers via Portal
Numerical Datasets	HEPData	DOI → Direct download
Metadata	InvenioRDM + databases	Portal discovery
Workflows	Containers, REANA, documentation	Executable via REANA
Documentation	InvenioRDM + Static HTML	AI search → Links
Code	GitHub & Gitea @ BNL	Versioned, linked, AI-indexed for reuse

Our Vision - End-to-End Reproduction

```
$ reproduce Figure 3 from STAR's jet suppression paper (Au+Au 200 GeV, 2007)
```

User Workflow

1. **Access Portal**: Start at central RHIC Data Portal with single sign-on authentication
2. **Search**: Enter "*STAR jet suppression 200 GeV Au-Au collisions 2007*"
3. **Discovery**: The system matches the paper, datasets, and analysis environment
4. **One-Click Access**: Follow direct links to analysis-ready datasets
5. **Launch Environment**: Open a preserved Docker container with exact software versions
6. **Execute Workflow**: Run documented workflow from containerized environment
7. **Verify Results**: Compare generated plots with original published Figure 3

Key Benefits

- No installation needed: analysis runs in preserved containers
- Full provenance: software, parameters, and data exactly match original work
- Easy to use: AI guidance lowers the barrier for non-experts

Working Prototypes

ChatBot System:

- Built upon ChatSTAR prototype
- Supports natural language queries across experimental documentation
- Provides context-aware responses from preserved knowledge base

InvenioRDM Platform:

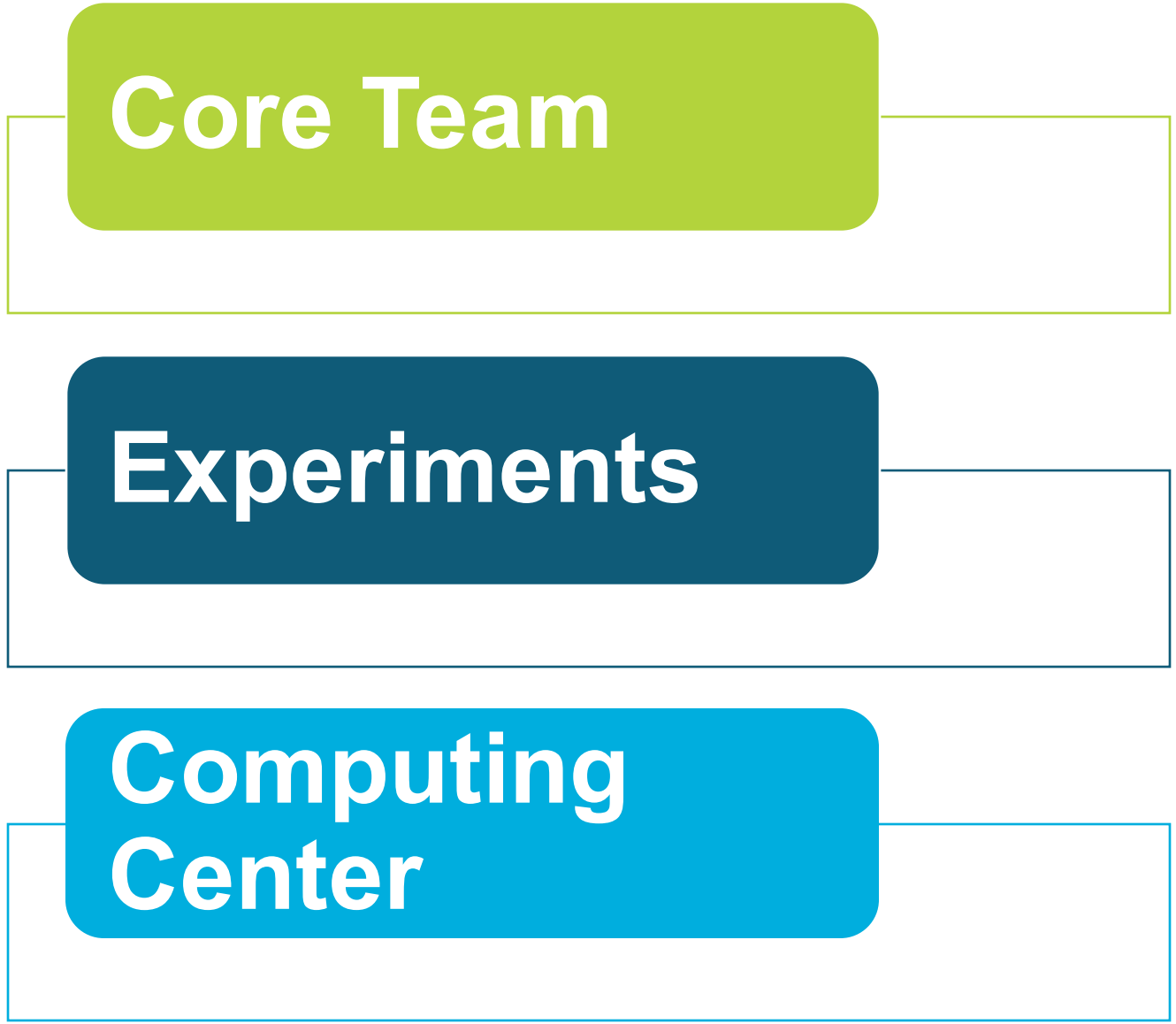
- Digital repository platform for data and publication management
- DOIs from OSTI (DOEs Office of Scientific and Technical Information) assignment and metadata handling capabilities

OpenData Portal:

- Public access interface for selected datasets
- Provide access to educational resources and documentation
- Demonstration of FAIR principles implementation
- Working on multi-level access with CERN
- Plan to connect to QuarkNet (HEP initiated) Masterclasses (BNL is one of the 50 QuarkNet centers)

Staffing and Organization

Collaborative Implementation Approach



People, Process, and Tools

- **Curators** within collaborations ensure that only well-documented, trusted data enters the system.
- **Data, software, and publications** are tightly linked to ensure reproducibility
- Each experiment maintains **responsibility** for validating and curating its preserved content
- Clearly defined **access levels** (collaborator, public, controlled) ensure users understand what is available and under what conditions.
- **A search engine** with natural language support helps users find relevant documentation and datasets

Success requires active collaboration, participation, and institutional support during Phase I.

Staffing requirements

- Realistic Implementation in two phases
  - Phase I (Years 1-5): 6.5 FTE
  - Phase II (Years 6+): 2.1 FTE

Role	Phase I FTE	Phase II FTE
Core Development Team	5.5	1.7
Experiment Coordinators	1.0	0.3
Computing Support	(0.3)	0.15
Total	6.5	2.15

Phase I Focus: System development, knowledge capture, infrastructure
Phase II Focus: Operations, maintenance, user support

Core Team Roles

- Heavier staffing in Phase I to support system design and implementation
- Leaner Phase II team focused on long-term access and support
- Positions are supported through DAPP

Role	Phase I FTE	Phase II FTE
DAPP manager	1.0	0.5
Software & Workflow	1.0	0.3
DAP Portal Development, Services Integration, and Administration	1.25	0.4
AI Integration	2.0	0.4
Documentation & QA	0.25	0.1
Total Core Team	**5.5**	**1.7**

Core Team:

- **Management**: DAPP Manager (oversight)
- **Technical**: Software preservation & containerization
- **Platform**: Portal development & system integration
- **AI Systems**: Knowledge extraction & semantic search
- **Quality**: Documentation & validation

Experiment Liaison

Preservation Coordinator (DAPP-funded) one in each experiment

Role	Phase I FTE	Phase II FTE
Preservation Coordinator (per experiment)	0.3	0.15
Total Experiment Support	1.0	0.3

Roles:

- Liaison between each experiment and the central team
- Ensures consistency and alignment of preservation efforts
- Supports practical implementation of DAPP within each collaboration

Brookhaven National Laboratory

Computing Data Center Support

- Dedicated DAPP-support roles at SCDF
- General computing services are also provided by SCDF
- Both are supported through RHIC operation

Role	Phase I FTE	Phase II FTE
Technology Watch Analyst	0.2	0.1
Computing Center Liaison	0.1	0.05
Total Computing Support	0.3*	0.15

*Supported by RHIC operation during Phase I
For Phase II, funding clarification needed

Roles:
Technology Watch: Monitor trends, plan migrations
Infrastructure Liaison: Coordinate with SCDF systems

Experiment Engagement Challenge

What's Required from Experiments

- **Scientific Validation**: Expert review and annotation of preserved records
- **Code Preservation**: Containerized workflows and execution documentation
- **Active Participation**: Training, documentation, and community engagement

Major Challenges

- Limited availability of dedicated effort at institutions
- Need active participation during Phase I while domain experts are available
- Institutional support needed, cannot rely on volunteers alone

Success Factor

- **Integration into publication workflows:** making preservation a standard part of the research process rather than an add-on

Hardware requirements

Hardware Resources

Hot Storage (~30 PB estimate):

- Fast disk-based storage for frequently accessed analysis objects
- Immediate access with integrity checks and snapshots

Cold Storage (~700 PB estimate):

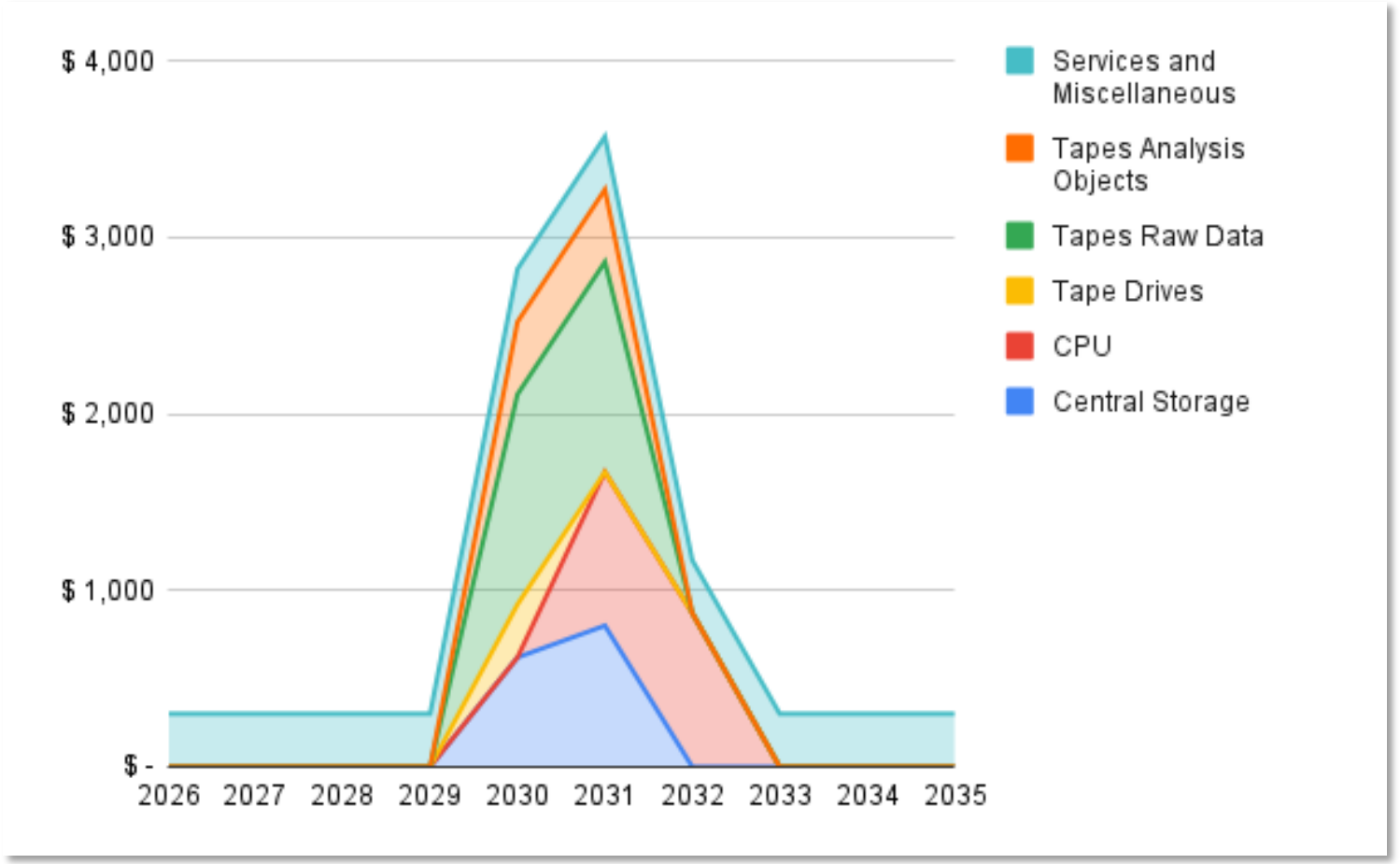
- Current libraries use LTO-8/LTO-9, Planned upgrade to LTO-11 after Phase I
- Automated migration between tape and disk via data carousel software

Computing

- Scaled to twice the current STAR capacity for analysis workloads (estimate)
- Opportunistic computing through SCDF and ASCR facilities

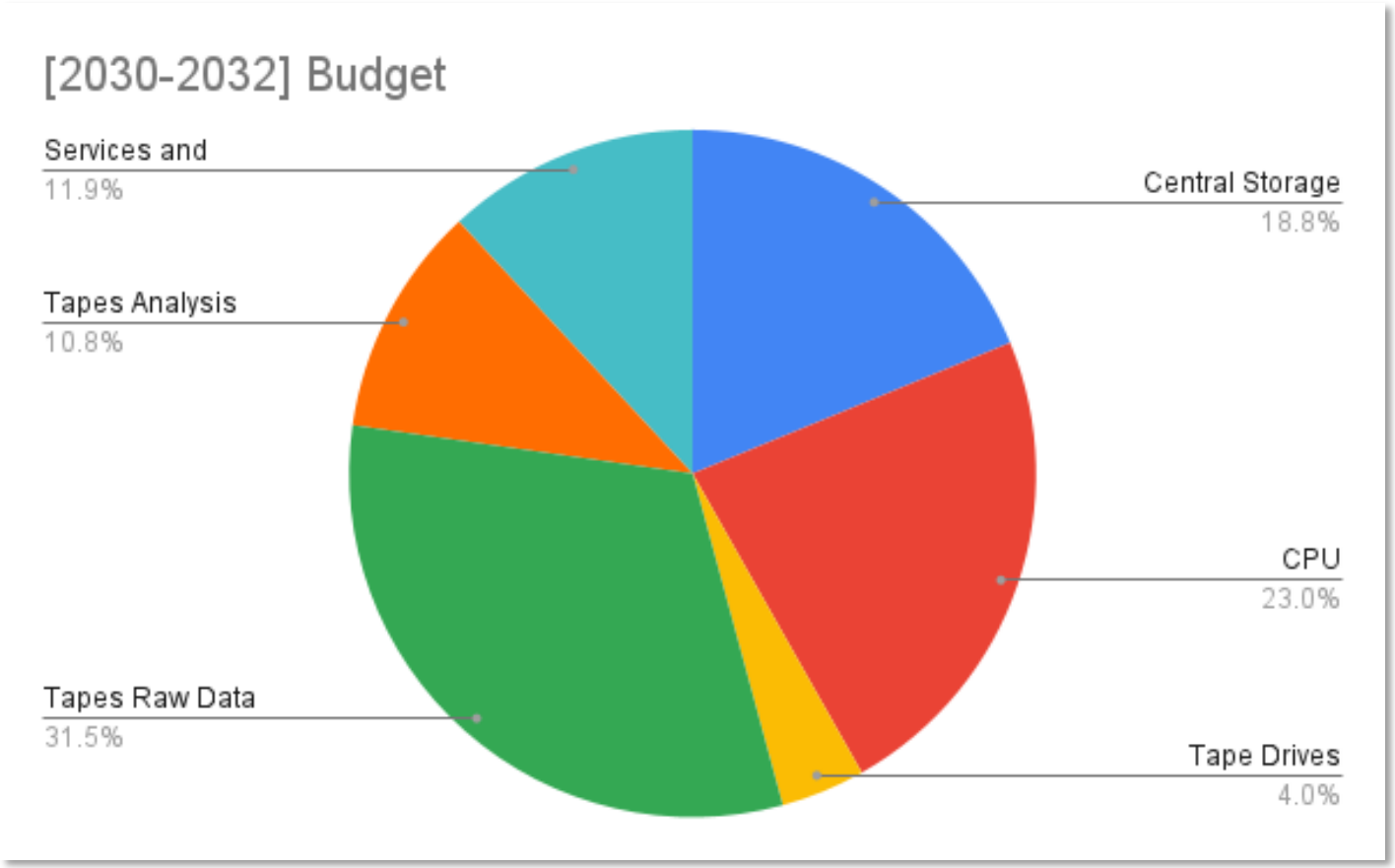
DAPP estimated hardware budget

- Major spike in 2030-3032, peak at $3.5M/y
- With this budget basic DAP infrastructure (CPU, Disk, Tape) is setup for the RHIC experiments
- First hardware refresh (storage) targeted in 2036



[2030-2032] hardware budget distribution

- Total Hardware Investment: ~$7.6M over 3 years
- Tape media ~ 40%



Implementation Timeline Overview

2026–2029: Foundation Phase

- Deploy core infrastructure: portal, storage systems, containerized environments
- Begin systematic data inventory and ingest high-priority datasets
- Develop and test AI-guided discovery tools

2030–2032: Production Phase & Infrastructure Refresh

- Major hardware investments (~$7M)
- Migrate data to LTO-11 tape and upgrade systems
- Build user community and collect feedback
- Launch production AI tools and user interfaces
- Complete ingestion of ~120 PB of analysis-ready data
- Transition to Phase II operations

2032 and Beyond: Sustainable Operations

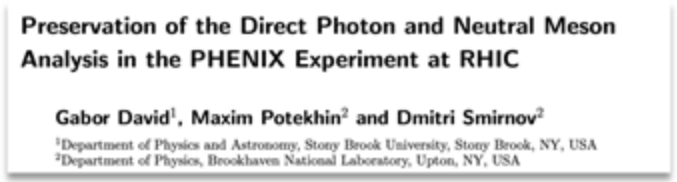
- Lean team focused on system upkeep, user support, and technology updates
- Ongoing data integrity checks and quality assurance
- Carry out scheduled hardware refresh cycles

Dependencies and Risks

Critical Dependencies

- **DOE Funding Continuity**: Sustained support through both Phases
- DOE **Office of Scientific and Technical Information** (OSTI): persistent identifiers and metadata services
- **ITD Infrastructure Services**: Core IT support and systems
- **SCDF Infrastructure**: Computing center support, storage systems, and maintenance
- **Computing and Data Sciences** (CDS) Directorate**:** AI expertise and specialized hardware

Building on Strong Foundations


Preservation of the Direct Photon and Neutral Meson Analysis in the PHENIX Experiment at RHIC

Gabor David[1], Maxim Potekhin[2] and Dmitri Smirnov[2]

[1]Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY, USA
[2]Department of Physics, Brookhaven National Laboratory, Upton, NY, USA

Experiment-Led Initiatives

- **PHENIX**: Static Web repository; containerized analysis, REANA at BNL
- **STAR**: Experience with containers, ChatSTAR AI prototype
- **sPHENIX & ePIC:** InvenioRDM
- **Publication practice**: Analysis must be reproducible and independently verifiable by non-team members

RHIC DAPP Approach:

- Develop **common, flexible solutions** adapted to each experiment's maturity
- **RHIC DAPP Roundtables** (est. 2024) foster coordination & best practices
- **Active prototyping**
- Developed on a lab-supported program

Prototyping & Engagement: Direct collaboration with PHENIX, sPHENIX and STAR teams to define features and validation of approaches

Technology Evolution Planning

Technology Watch:

- Track storage, compute, and container trends
- Plan tape refreshes and platform shifts (e.g., ARM/x86)
- Review APIs, standards, and security protocols

Platform Independence:

- Multi-architecture containers (x86 & ARM)
- Open standards: ROOT, CWL, Docker
- Modular systems avoid vendor lock-in

Quality Assurance

Expert Validation:

- Experiments select and validate their records
- Only documented, expert-reviewed record enters the archive

Long-Term Integrity:

- Automated checks detect corruption or degradation
- Complete provenance tracking from raw data to final results
- Regular audits ensure verify continued accessibility and functionality

Standards-Based Preservation:

- Documented workflows and metadata requirements
- Version control and change tracking
- Format migration planning for long-term access

Long-Term Usability Requirements

Technical Sustainability:

- Use sustainable, formats: (ROOT, text, JSON)
- Avoid locked software or platforms
- Plan for migration: tapes, containers, and storage must evolve

Knowledge Preservation:

- Preserve context: document how the data were used and the reasoning behind analyses
- Capture expert knowledge while it's still exist
- Ensure external users can discover and use data effectively
- Maintain integrity: perform regular unit testing

Overview of Risk Management

16 Specific Risks Identified Across 4 Categories:

1. **Data Quality and Processing** - Hardware failures, data corruption, validation challenges
2. **Technology and Infrastructure** - Platform obsolescence, architecture changes, storage evolution
3. **Institutional and Funding** - Budget discontinuation, priority changes, personnel loss
4. **User Adoption and Community** - Engagement, sustainability, educational development

Risk Examples

Data Quality and Processing:

- Multiple reprocessing cycles may be needed for final datasets
- Data corruption during tape migration - Low probability, high impact

Technology and Infrastructure:

- Disk storage failures likely during preservation period
- Potential obsolescence of container technology

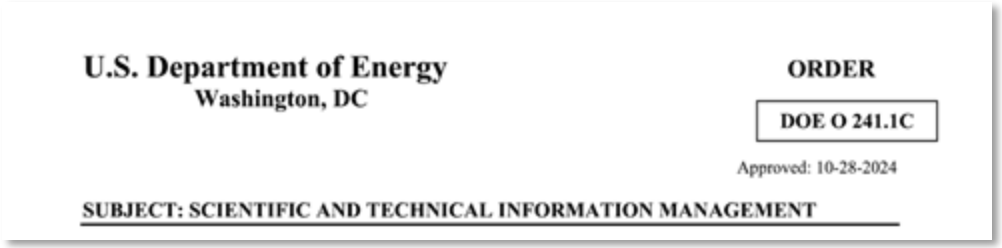
Institutional and Funding:

- Possible discontinuation of Phase II funding
- Loss of key personnel affecting operations

DOE requirements & best practices

Regulatory Alignment

- **DOE Order 241.1C:** Scientific and Technical Information Management compliance
  - Will work with OSTI contacts on this
  - 241.1C explained here
- Office of Science and Technology Policy (**OSTP**) Nelson **Memorandum:** Public access to federally funded research data
  - DOI assignment for public datasets and documentation
- **Cybersecurity:** BNL cybersecurity protocols and federal compliance



U.S. Department of Energy
Washington, DC

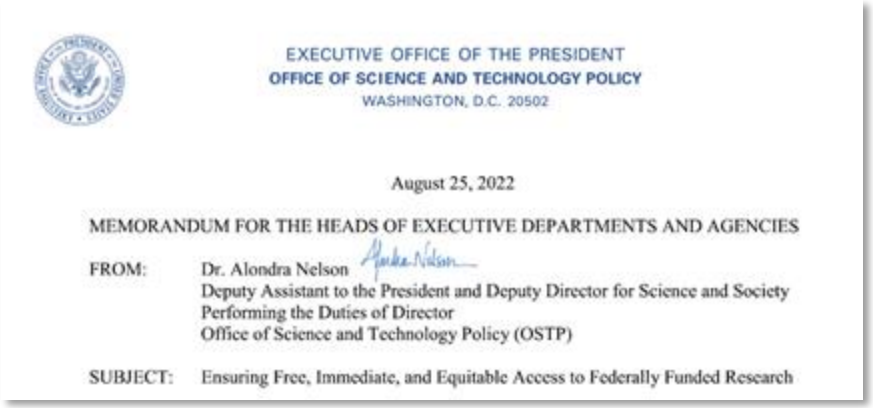
ORDER

DOE O 241.1C

Approved: 10-28-2024

SUBJECT: SCIENTIFIC AND TECHNICAL INFORMATION MANAGEMENT

In summary, DOE Order 241.1C mandates that all scientific and technical results from DOE-funded research be submitted to the DOE Office of Scientific and Technical Information (OSTI) so they can be preserved, shared, and made publicly accessible when appropriate.



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

In short, the OSTP Nelson Memorandum requires all federally funded research publications and data to be made freely available to the public upon publication.

Standards and Ethical Framework

International Data Standards:

- Implement FAIR principles to ensure data are Findable, Accessible, Interoperable, and Reusable
- Pursuing CoreTrustSeal certification, following the OAIS (Open Archival Information System) framework for Open Archival Information System

Compliance Requirements

- Coordination with BNL Legal Affairs on export control and intellectual property matters, regular audits
- Clear documentation of data provenance and usage rights

Goal: Build a trusted, compliant preservation system for long-term scientific use

Impact and Recommendations

Cost-Benefit Context

Investment Perspective:

- DAPP Preservation Cost: ~$10M over 10 years
- Less than 0.05% of total investment secures long-term scientific value

Alternative Costs:

- Full data reprocessing alternative: tens of millions
- Commercial cloud storage alone: ~$30M per year

DAPP approach: Cost-effective, balanced solution

What DAPP Can and Cannot Guarantee

What DAPP Can Achieve

- Preserve analysis-capable datasets for continued scientific research
- Reproducibility for key published RHIC analysis
- Support educational use of real experimental data
- Extend scientific return on RHIC investment by 10+ years

What DAPP Cannot Guarantee

- Reconstruction of all preserved data in a timely manner
- Long-term preservation without continued institutional support
- Seamless migration to alternative institutions if institutional support ends
- Preservation of all analyses, especially undocumented legacy work

Success Metrics

Scientific & Educational Impact

- Number of new papers citing preserved data
- Portal, Dataset and DOI use

User Engagement

- Number of new users per year
- Number of success in reproducing analyses
- Satisfaction scores

Preservation Quality

- Data integrity
- Full coverage of data per experience
- Complete analysis setups saved

Operational Performance

- Costs in line with projections
- DOE and FAIR compliance maintained

DAPP Fully Addresses All Committee Questions (1–3)

Q1: Scientific Value: Does the DAPP effectively identify and preserve the most valuable scientific assets and legacy from the RHIC experiments?

DAPP preserves what matters most: analysis-ready data, working code, and expert knowledge from 25 years of RHIC research. This ensures continued access to experimental data that can't be recreated elsewhere.

Q2: Infrastructure Capability: Will the proposed infrastructure enable both verification of published results and new analyses by external researchers?

Yes. Researchers can reproduce published results and run new analyses through web browsers, with no software installation needed. Everything operates in containers using the exact software versions from the original work.

Q3: Long-Term Usability: Are proposed data curation practices sufficient to ensure long-term usability and discoverability of RHIC data?

DAPP adheres to FAIR data principles and incorporates AI search tools to assist users in finding relevant data and documentation. The system works even after the original RHIC experts retire.

DAPP Fully Addresses All Committee Questions (4–5)

Q4: Resource Realism: Are the proposed FTE allocations and infrastructure requirements realistic for both the initial and sustained implementation phases?

Yes. The plan uses a realistic, phased resource model that begins with expanded staffing to develop the system and then shifts to lean operations for long-term maintenance. It builds on BNL's existing computing infrastructure.

Q5: Risk Management: Has the plan identified major risks and outlined suitable mitigation strategies?

The plan identifies 16 specific risks along with response strategies. Key components such as the AI search system and data repositories are already functional prototypes, which reduces technical risk.

DAPP Summary

Resource Requirements

- **Phase I** (Years 1–5): 6.5 FTE
- **Phase II** (Years 6+): 2.1 FTE for sustainable operations
- **Hardware Budget**: ~$7 M in 2030–2032

Critical Dependencies

- Experiment engagement depends on institutional funding to provide dedicated preservation roles within each collaboration.
- ITD infrastructure, CDS AI expertise, and SCDF computing and storage services
- Ongoing support from DOE needed during both implementation and operations.

DAPP provides a cost-efficient, two-stage framework to protect RHIC's scientific legacy.

DAPP: Securing RHIC's Scientific Legacy

The opportunity before expertise disperses as RHIC operations end

Proposed Solution

- Prototypes validated with experiments
- AI systems will index 25 years of documentation
- InvenioRDM platform deployed for record and metadata management

Expected Impact

- Extend RHIC’s scientific output by a decade or more
- Enable new discoveries as theory advances
- Provide real data for education and training
- DAPP tools and workflows will benefit future projects like EIC
- Position BNL as a model for future preservation efforts

DAPP is Ready to Proceed

Resource Plan

- Realistic two-phase staffing: 6.5 FTE → 2.1 FTE transition model
- Hardware investment: $7.6M over 3 years (2030-2032)

Risk Management in place

- 16 specific risks with mitigation strategies
- Plan for scaling down services if funding is limited
- Technology evolution planning with annual reviews

Success Depends On

- DOE funding sustained through both phases
- Strong engagement from experiments during Phase I
- Institutional support to fund preservation roles within experiments
- Continued BNL infrastructure and service support

Framework for Success

RHIC's legacy can be preserved; success will depend on:

- Members of collaborations actively contributing
- Support from institutions for key roles within experiments
- Early coordination among stakeholders to ensure a smooth transition from development to long-term operations
- A long-term commitment to keeping data accessible and usable

It is time to ensure RHIC's legacy continues to deliver scientific value

Thank you!