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.

- TABLE OF CONTENTS

1 Strategic Vision

1.1 RHIC's Scientific Legacy

Since 2000, RHIC has been a world-leading nuclear physics facility, having discovered quark-gluon plasma and produced over 600 publications. The preserved datasets capture unique energy regimes and collision systems, providing research opportunities that are unavailable at current or planned facilities.

The Preservation Challenge

Preserving RHIC data presents significant challenges due to the large volume of data, experimental complexity, and limited resources.

- **Scale and Complexity:** Nearly 500 petabytes of data from three experiments, PHENIX, sPHENIX, and STAR, each featuring unique detector systems and independently developed data formats, accumulated over 25 years. The upcoming 2025 run introduces challenges, potentially doubling the data volume collected by RHIC over its history..
- **Knowledge Transfer:** Future users will likely lack the deep experimental expertise and specific experience of the the current scientists.
- **Resource Constraints:** Preservation must proceed with a shrinking workforce and aging infrastructure approaching retirement.

Strategic Objectives

The goals include enabling continued scientific discovery that maximizes DOE's investment, ensuring data integrity for result verification, protecting significant scientific and financial assets, building reliable infrastructure that meets federal requirements and aims for CoreTrustSeal certification, and supporting a broad user base collaborators, new researchers, educators, and other stakeholders.

1.2 Implementation Strategy

Two-Phase Approach

The preservation plan follows a two-phase model adaptable to evolving resources and infrastructure.

Phase I: Active Implementation (Years 1-5) occurs while collaborations remain engaged and the current computing infrastructure is available. Divided into Foundation Building (Years 1-2) and Full-Scale Implementation (Years 3-5).

Phase II: Long-term Stewardship (Year 6+) transitions to sustainable operations with reduced resources while retaining core capabilities.

Community Collaboration and Best Practices

RHIC DAPP collaborates closely with the Data Preservation in High Energy Physics (DPHEP) group [13], leveraging shared tools, standards, and community expertise. Partnerships with CERN, DESY, KEK, universities, and cloud providers will expand the program's reach and reinforce DOE's leadership in scientific data preservation. Engagement with the ICFA Data Lifecycle Panel [14] connects RHIC efforts to global initiatives that manage data from acquisition through reuse, emphasizing open science and FAIR(Findability, Accessibility, Interoperability, and Reusability) principles [5]. This collaboration enhances international coordination and promotes recognition for data lifecycle work within the research community.

Flexible Timeline

The transition timeline will remain flexible, depending on factors such as the completion of key data reprocessing, especially the 2025 final Au-Au run, the readiness of preservation infrastructure, shrinking active collaborations, and the retirement of legacy hardware. Annual reviews will evaluate these elements to decide the best time for phase transitions.

2 Preservation Approach

The preservation effort covers all components necessary to ensure scientific reproducibility. Priority is placed on Analysis Objects, which will be readily usable for verification and new studies. While the Raw Data archive will be retained, its use will be constrained by limited CPU resources and tape access capacity, making large-scale reprocessing infeasible in the long term.

The estimated data volumes for the three major experiments are shown in the table below. The current planning estimates a total of approximately 750 PB of data. The size listed for Analysis Objects reflects a single version of processing.

Table 1: *Estimated data volumes by type of data and experiment*

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

Software and analysis workflows are preserved using container technology. Accompanying documentation, such as technical notes and contextual details, will be kept alongside metadata that describes experimental conditions, detector setups, and data processing parameters.

2.1 Practical Preservation Model

In line with the standards of DPHEP, the approach strikes a balance between scientific needs and resource realities. We prioritize maintaining access to processed physics data, analysis software, and documentation, enabling future researchers to perform analyses without the need to reprocess raw data. Both simulation and reconstruction software will be preserved.

While all raw data will be archived, large-scale reprocessing is impractical due to the limited computing resources available. Achieving complete reprocessing capability would require sustaining an expensive computing infrastructure.

Additionally, analysis environments are preserved, including container images and analysis notebooks, which capture complete software stacks to ensure they remain functional and

reproducible long after their original development. These innovations strengthen long-term access to both knowledge and tools.

2.2 Global Reprocessing

The plan includes a final, comprehensive reprocessing of sPHENIX and STAR data to produce consistent and well-documented datasets. By utilizing the latest software versions and accumulated experimental knowledge, this process will not only eliminate variations from past run periods but also significantly extend the scientific lifetime and usability of the produced data. This expertise will also support more precise provenance tracking by defining definitive data versions and documenting workflows from raw data through analysis objects. It will also potentially ease the production of special datasets for educational purposes. The timing and extent of this reprocessing will depend on the volume of the 2025 Au-Au data, which may necessitate multiple processing passes and consume substantial computing resources.

2.3 AI-Enhanced Preservation

RHIC DAPP uses AI-driven search technology to provide natural language queries across 25 years of experimental documentation. This makes it easier and faster for users to find relevant information within a large collection of technical documents, improving both accessibility and efficiency.

Building on the successful STAR prototype, the system has been trained on thousands of internal RHIC documents. Instead of reading through lengthy technical notes, users can ask questions like "How do I apply track quality cuts?" and get relevant code snippets and parameters. The role of each experiment is critical; they organize, annotate, and validate their records to ensure that only accurate, well-curated, and current information is included. Since this knowledge cannot be transferred automatically, AI Integration Specialists on the DAPP team work directly with experiments to help capture expert knowledge and translate it into formats the AI system can use. This collaborative process ensures that users receive reliable, experiment-specific insights without needing to navigate through irrelevant or outdated content.

The AI-enhanced preservation in RHIC DAPP extends beyond Retrieval-Augmented Generation (RAG) [11] and Model Context Protocol (MCP) [12], which are currently part of the prototype. It also integrates advanced natural language processing and containerized workflow preservation. Custom-trained language models understand nuclear physics terminology, experimental contexts, and analysis methods. This allows researchers to access decades of knowledge with straightforward, plain-English queries, without needing expertise in specialized databases.

Additionally, AI techniques can transform collections of scripts and code into clean, well-organized workflows packaged as container images. This ensures complex analysis processes are captured accurately and remain easy to reproduce.

The AI domain is evolving rapidly, with ongoing advances in tools, methods, and standards. RHIC DAPP actively monitors and adapts to these developments to ensure its approach remains aligned with emerging paradigms and best practices. Staying current in this dynamic field is essential to maintain scientific relevance and technical robustness.

More details are provided in Appendix A.

3 TECHNOLOGY INFRASTRUCTURE

3.1 SERVICE ARCHITECTURE

The preservation system operates within the SCDF's containerized infrastructure, currently built on OpenShift [15], the Kubernetes-based platform developed that enables to build, deploy, and manage containerized applications at scale. It supports a diverse user base, including collaborators, new researchers, and educators, through dedicated access interfaces. Central to this ecosystem is the RHIC Data Portal, which manages user authentication, search, and content navigation. To enhance usability, a prototype AI assistant is in development. It will guide users in locating and understanding preserved materials based on their roles, background, and specific needs. This functionality introduces the requirement for a robust Role and Identity Management System (RDMS) to manage authentication, access permissions, and user profiles.

Several dedicated portals support different aspects of data preservation and access. The OpenData Portal [6] provides public datasets and educational resources. Publications and documentation are managed through an InvenioRDM repository [8] with DOI support. Preserved workflows can be executed via the REANA platform [9], the open-source platform developed at CERN that enables reusable and reproducible research through containerized workflows. Interactive Jupyter environments facilitate data exploration, while Git repositories manage code and handle version control.

3.2 STORAGE STRATEGY

The storage architecture employs a tiered model that balances quick access and long-term preservation. Hot storage relies on fast and cheap storage for about 30 PB of frequently accessed analysis objects, providing immediate access with integrity checks and snapshots.

Cold storage utilizes BNL's tape libraries using LTO-8/LTO-9 technology, with a planned upgrade to LTO-11 after Phase I. BNL's data carousel system automates the intelligent migration between tape and disk, using predictive algorithms that will be developed to pre-stage frequently requested data based on user access patterns.

3.3 COMPUTING RESOURCES

In Phase I, the plan continues to rely on the existing RHIC computing infrastructure to support ongoing data analysis and reprocessing (see Appendix B). In Phase II, computing resources are scaled down to about 60% of STAR's current capacity, based on past usage showing that data analysis typically accounts for that portion of total computing needs. Additional capacity can be provided by opportunistic computing at BNL SCDF and collaborations with DOE ASCR facilities as needed.

3.4 USER EXPERIENCE AND ACCESS

The preservation system defines three levels of data access designed to balance openness, scientific reliability, and data integrity.

- Collaboration-restricted access is available to verified experiment members, who can access protected datasets through existing authentication systems.
- Public access applies to selected datasets associated with peer-reviewed publications. These are made openly available, accompanied by appropriate documentation, DOIs, and metadata.
- Controlled access applies to datasets that require specialized domain-specific expertise to be interpreted correctly. Access is granted on a case-by-case basis to avoid misinterpretation or misuse.

The RHIC Data Preservation Portal will be designed as a unified, AI-assisted interface that connects users to preserved data, workflows, documentation, and metadata across distributed systems ensuring long-term sustainability.

4 GOVERNANCE AND SUSTAINABILITY

4.1 GOVERNANCE STRUCTURE

The RHIC DAPP operates within a streamlined governance model designed to facilitate effective decision-making while ensuring accountability to stakeholders and institutional leadership.

The Steering Committee provides strategic oversight, consisting of representatives from each RHIC experiment, NPP Directorate leadership, and external experts. It meets twice a year during active implementation and once a year during the maintenance phase.

The Technical Working Group advises the Steering Committee and the Implementation Team on infrastructure options, preservation standards, and community best practices. It includes representatives from BNL ITD, SCDF, and DAPP technical leads.

The Implementation Team manages daily operations under the leadership of the DAPP manager, making operational decisions within the strategic framework and reporting progress to the Steering Committee.

4.2 INSTITUTIONAL AND OTHER DEPENDENCIES

The RHIC DAPP builds on BNL's established computing infrastructure and expertise in scientific data management, supported by both the RHIC operations program and institutional services. This foundation facilitates long-term preservation and cost-effective operations through shared institutional resources.

DAPP depends on two main categories of support: services from BNL's IT Division and RHIC-funded services managed by the Scientific Data and Computing Facility (SCDF). Continued institutional backing is an essential requirement that must be addressed proactively to sustain the program over time.

BNL's IT Division provides institutional services that support secure collaboration and infrastructure:

- 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

The RHIC operations program funds essential computing and data services that are managed operationally by the Scientific Computing and Data Facilities (SCDF). The SCDF provides technical expertise, manages infrastructure, and day-to-day operations for these RHIC-funded services, reducing costs and complexity while ensuring alignment with established technical standards and leveraging institutional operational capabilities.

- 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

The BNL Computing and Data Science Directorate provides additional specialized services, including access to state-of-the-art computing resources specifically designed for AI applications and related technical expertise.

During Phase I, all service dependencies will be documented, and contingency plans will be prepared. This phase will also demonstrate DAPP's relevance to both the RHIC program and the broader BNL mission. Before entering Phase II, the team will work with stakeholders to ensure continuity of critical services or develop alternatives, such as cloud solutions or external partnerships.

5 Resource Requirements

5.1 Staffing Overview

The staffing outlined below covers only the centralized RHIC effort required to implement the DAP Plan. It does not include the additional effort needed within the RHIC collaborations themselves. The plan assumes that participating institutions will be able to support their own contributions to data preservation. However, that assumption may not hold. Many groups that have

historically contributed essential analysis and software expertise are under serious funding pressure. Without targeted support for their continued involvement, this plan cannot succeed.

The preservation effort proposes a team structure that includes both dedicated specialists and part-time contributors drawn from the experiments and the computing center. Team roles are summarized in the following table and described in detail in the next section and Appendix C.

Role	Phase I FTE	Phase II FTE
Core DAPP Team	5.5	1.7
Preservation Coordinators	1.0	0.3
Computing Support *(Provided by RHIC Operration during Phase I)*	*(0.3)*	0.15
Total	***6.5***	***2.15***

CORE DAPP TEAM

The central preservation team provides dedicated expertise in building and managing the preservation infrastructure. Its size and composition reflect the scale and technical complexity of the RHIC data, as well as its long-term preservation goals. Details of the DAPP-supported roles are listed below.

Role	Phase I FTE	Phase II FTE	Primary Responsibilities
DAPP manager	1.0	0.5	Leads overall coordination, strategic planning, stakeholder engagement, and governance alignment.
Software & Workflow	1.0	0.3	Preserves analysis software, builds containerized environments, and translates workflows into reusable formats. Involves close coordination with experiments to align with actual analysis practices and capture essential processing context.
DAPP Portal Development, Services Integration, and Administration	1.25	0.4	Develops and maintains the Invenio-based OpenData and image repository platform, integrating metadata pipelines and ensuring access, reuse, and interoperability through CI/CD workflows and standards-compliant services.
AI Integration	2.0	0.4	Develops and maintains AI systems for semantic search, knowledge extraction, and integration with large language models (LLMs). Collaborates with experiments to transfer analysis knowledge and domain expertise into AI systems. Responsibilities include translating scripts and codebases into long-term containers and workflows, building pipelines to index content from wikis, software, metadata, and documentation, capturing expert knowledge across all three experiments, and supporting a modular design to accommodate future LLM updates.
Documentation & QA	0.25	0.1	Writes clear documentation, training materials, checks that processes and outputs meet quality standards.
Total Core Team	**5.5**	**1.7**	

Table 1: Core DAPP Team Structure

Essential Functions Within Experiments

While the central team builds infrastructure, the experiments are expected to actively contribute, as they best understand their data formats, software, and scientific needs. Each experiment should appoint collaborators to help define and implement preservation activities.

These roles are described in Appendix C. It is important that each collaboration actively integrates DAP into its publication and analysis process. This requires that efforts and the support thereof

within the collaborations are considered essential to the success of the plans proposed in this document. These contributions involve dedicated effort and cannot rely solely on volunteer work, even when integrated into existing workflows.

The Preservation Coordinator, funded by RHIC DAPP, plays a special and important role. As the liaison between each experiment and the central team, the coordinator helps ensure consistency and alignment across all experiments and supports the practical implementation of the DAPP within each collaboration.

The FTE numbers in the table reflect the part of this role funded through the central DAPP. The actual effort needed may vary by experiment, but overall, about 1 FTE is allocated across the three experiments. For the plan to succeed, each experiment also needs collaborators involved in preservation tasks alongside the coordinator. That kind of engagement requires more than institutional endorsement. It relies on dedicated funding to make the necessary effort possible. Without that support, even committed groups may not be in a position to contribute. The coordinator alone cannot carry the effort alone.

Table 2: Experiment-Specific Roles

Role	Phase I FTE	Phase II FTE	Primary Focus
Preservation Coordinator (per experiment)	0.25 - 0.3	0.15	Represent experiment needs, coordinate with the DAPP team, and define preservation goals
Total Experiment Support	**1.0**	**0.3**	

Computing Data Center Support

The preservation effort relies on services provided by SCDF staff in two categories: dedicated DAPP-support roles and general computing services that maintain and operate the underlying infrastructure. During Phase I, both categories of support will be provided through the RHIC operations program, including the dedicated DAPP-support roles at SCDF. Additional details on these roles are provided in Table 3.

Table 3: Computing Center Support

Role	Phase I FTE	Phase II FTE	Primary Responsibilities
Technology Watch Analyst	0.2	0.1	Monitor technological trends, assess emerging technologies, and prepare annual reports

Computing Center Liaison	0.1	0.05	Coordinate between the DAPP team and the computing infrastructure
Total Computing Support	**0.3***	**0.15**	**Supported by RHIC operation during Phase I*

As the RHIC DAPP transitions into Phase II, clarification will be needed regarding ongoing funding for these positions, particularly as RHIC operations funding winds down.

Transition Planning

In Phase I, these general services will remain under RHIC operation. To prepare for Phase II, a transition plan will be developed that covers funding, staffing, and sustainable infrastructure requirements. This includes identifying any hardware or services that need upgrades or transfers before the conclusion of the RHIC operation program.

5.2 Resource Planning

References to computing and storage in this document refer to resources provided through the RHIC operation program and managed by SCDF. Annual planning will be coordinated jointly between RHIC and SCDF to ensure accurate forecasting, cost control, and continuity of operations through both phases.

5.3 Infrastructure Requirements

Current CPU and disk resources, sized for the sPHENIX streaming model, should easily support offline processing and analysis for years.

The storage system uses a tiered model, with approximately 30 PB of disk-based hot storage for fast access to analysis-ready data, and 400–600 PB of cold storage on tape for archiving raw data. A migration to LTO-11 tape technology, expected to be released in 2027, is planned to be completed by the end of Phase I. Computing needs will be initially met using existing RHIC resources, then scaled to approximately 60% of current STAR computing capacity in Phase II, reflecting historical usage patterns during analysis periods. Details are provided in Appendix D.

5.4 Hardware Budget

The hardware-related budget is shown in Table 4 as a function of time. A noticeable increase in costs occurs around FY30–FY31, tied to the expected obsolescence of key computing infrastructure. Investments in tapes and drives, CPUs, and central storage begin in FY30 and peak the following year, reflecting the need to refresh aging systems to maintain reliable long-term access to the data.

Until then, the budget stays relatively flat, focused on core services and ongoing support. This includes a workflow submission platform, access to BNL's AI platforms managed by the Data and Computing Science Directorate, key software licenses and cloud-based services, as well as

dedicated hardware for the DAP Portal — the primary gateway to preserved RHIC data and documentation.

Aside from routine maintenance, no major hardware upgrades are planned during the first 5 to 7 years of Phase II.

Table 4: Annual Budget Projections

CATEGORY / [K$]	FY26	FY27	FY28	FY29	FY30	FY31	FY32
CENTRAL STORRAGE	$ -	$ -	$ -	$ -	$ 620	$ 800	$ -
CPU	$ -	$ -	$ -	$ -	$ -	$ 870	$ 870
TAPES AND DRIVES	$ -	$ -	$ -	$ -	$ 1,900	$ 1,600	$ -
SERVICES AND MISCELEANEOUS	$ 300	$ 300	$ 300	$ 300	$ 300	$ 300	$ 300
TOTAL	$ 300	$ 300	$ 300	$ 300	$ 2,820	$ 3,570	$ 1,170

6 RISK MANAGEMENT AND COMPLIANCE

6.1 KEY RISK CATEGORIES

Risks can be categorized into four broad areas: data quality, technology, institutional support, and community engagement. Risks related to technology include the failure of storage hardware and the potential obsolescence of container platforms. On the institutional side, Phase II funding remains uncertain, and future reductions in computing support at BNL might occur. Community engagement is also a concern, as participation may decline over time. Even as new facilities like the EIC become operational, RHIC data will continue to provide significant long-term scientific value.

Mitigation strategies are detailed in Appendix E and include flexible timelines, robust backup systems, and contingency plans for funding and infrastructure continuity.

6.2 COMPLIANCE FRAMEWORK

The plan complies with DOE Order 241.1C [1] and the OSTP Nelson Memorandum [2] by supporting public access to federally funded research, comprehensive data management, and submission of Scientific and Technical Information (STI), including DOI assignment for public datasets and documents. It also adheres to the FAIR principles [5], ensuring data remains Findable, Accessible, Interoperable, and Reusable. The team is pursuing CoreTrustSeal certification [4] to establish the repository as a reliable and credible preservation system. In coordination with BNL's

cybersecurity team, regular security audits and vulnerability checks are conducted to protect the data and meet federal compliance requirements.

7 Success Metrics and Sustainability

RHIC Data and Analysis Preservation is evaluated annually to ensure it remains useful, reliable, and aligned with evolving needs. Evaluations focus on five main areas. Completeness considers whether essential datasets, workflows, and documentation have been properly captured. Accessibility and usability examine whether researchers can locate and work with the preserved content effectively. Scientific impact looks at whether the data is leading to new publications, citations, or follow-up studies. Resource use is considered in relation to actual demand and long-term value. Policy compliance is checked against federal requirements, FAIR principles, and community standards. To ensure continued functionality, preserved workflows are also tested on a regular basis. These evaluations help guide improvements and make the case for sustained support.

Technology Watch Reports are published annually to track new developments in storage, repository systems, and computing. They help identify upgrade opportunities, flag potential risks from aging systems, and inform future infrastructure decisions.

8 Conclusions

RHIC DAPP safeguards data from over 600 publications and decades of DOE-supported research, representing several billion dollars in investment. It builds on established preservation methods from high-energy physics and incorporates AI tools to make RHIC data more accessible to future users. The plan is designed to be flexible, DOE-compliant, and realistic regarding its resource assumptions.

Its success will depend on strong institutional backing, engagement from the experiments, and sustained Phase I funding. Community use and contributions will be essential to ensure RHIC data continues to generate value, or both scientific discovery and training, for many years to come.

Appendix A DETAILED AI-ENHANCED PRESERVATION STRATEGY

RHIC DAPP implements an advanced AI-driven approach to preserving and accessing 25 years of experimental knowledge.

Semantic Knowledge Base: Utilizing Retrieval-Augmented Generation (RAG), it transforms extensive scientific documentation and software code into an intelligent, interactive resource that captures both data and tacit expertise.

Natural Language Understanding: Custom AI models interpret nuclear physics terms and contexts, enabling users to query the archive in plain language — in English and multiple other languages — without specialized expertise or database skills.

Model Flexibility and Evaluation: Through the Model Context Protocol (MCP), the system can integrate multiple large language models (LLMs), enabling comparative evaluation of outputs and easy substitution of AI components as technologies evolve.

Secure Access: Federated identity controls ensure collaboration documents are accessible only to authorized users, while public data remains open to all.

Predictive Data Staging: The system anticipates frequent data requests and preloads data from archives to reduce wait times.

Containerized Workflows: Complete analysis environments are preserved using container images, guaranteeing reproducibility decades later.

All AI and container tools will be open sourced for wider scientific use.

Key Features:

Knowledge Capture & Transfer: Builds on STAR prototypes to convert static documents into a dynamic AI-powered knowledge base, answering complex queries and generating relevant code snippets.

Automated Documentation: Extracts metadata and workflows from files and logs, explaining not just processes but the reasoning behind them. Supports modernization by translating legacy scripts into standardized workflows.

Enhanced Search & Recommendations: Semantic search enables natural language queries; recommendation engines suggest related resources, aiding researchers and newcomers alike.

Continuous Improvement: Regular updates, user feedback, and the integration of new AI technology ensure the system evolves without requiring full retraining.

Robust Deployment: Hosted at BNL, the system supports flexible integration of domain knowledge; for example, through the Model Context Protocol (MCP). Built-in Quality assurance includes expert review and citation of AI-generated responses.

Appendix B RHIC RESOURCES AND EVOLUTION OVER TIME

The RHIC program currently utilizes extensive computing resources at BNL. The CPU farm includes 150,000 job slots for approximately 2,300 kHS06, and the central storage capacity is about 150 PB, employing a combination of Lustre and dCache systems. CPU hardware typically lasts for 7 years, while disk storage generally lasts for 5 years before it needs replacement.

Since current CPU and disk resources are primarily provisioned to meet the demands of the sPHENIX streaming model, they are expected to comfortably support offline data processing and analysis for several years to come.

According to current budget assumptions, purchase dates, and equipment lifetimes, both CPU and disk resources are expected to decline significantly over time. By 2030, CPU availability for RHIC is projected to be around 50% of current levels. Disk storage is expected to experience a further reduction, decreasing to approximately 20% of its current capacity by 2030.

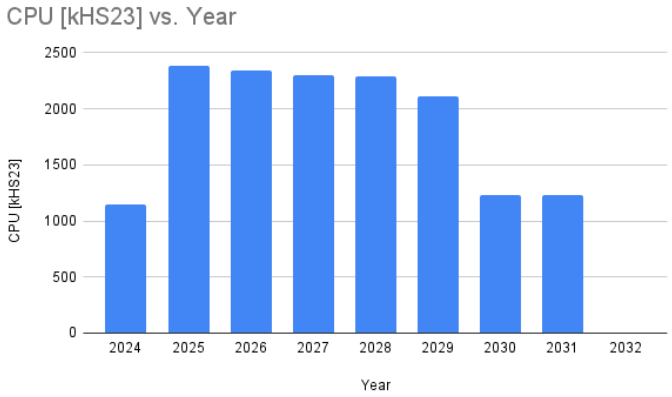


Figure 2: Evolution of the available CPU for the RHIC program

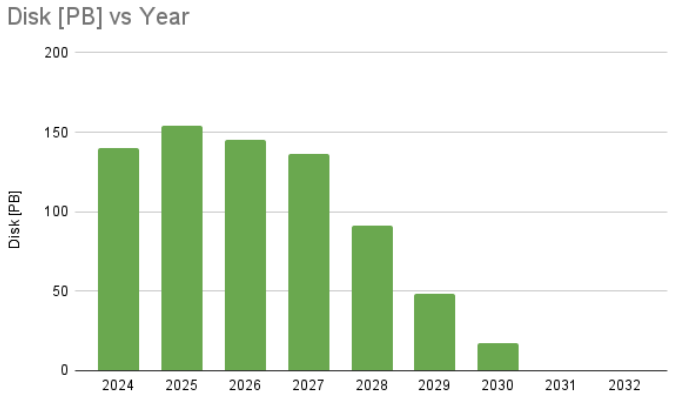


Figure 3: Evolution of the Disk storage available for RHIC

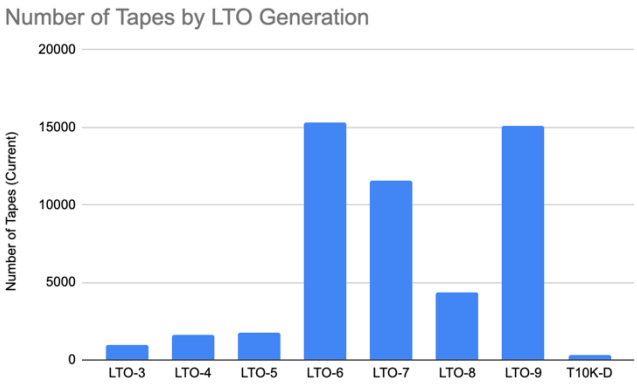


Figure 4: Distribution of the number of tapes used by RHIC per LTO type

For tape storage, the RHIC program uses seven tape libraries (5 Oracle and 3 IBM) with approximately 70,000 tape slots. The system uses LTO technology across multiple generations. Data gets migrated to newer tape media roughly every 2 LTO generations, with an average of 4 years between LTO generations. This means data moves from one LTO generation to another (called tape repacking) every 8 years.

During repacking, a second copy of the data is created if one doesn't already exist and the budget allows. Current plans call for LTO-8 tapes to be repacked onto LTO-10 media around 2028, with LTO-9 tapes moving to LTO-11 around 2030.

Appendix C STAFFING REQUIREMENTS

This appendix provides additional context on the staffing roles that support the RHIC Data and Analysis Preservation Plan (DAPP). Detailed role descriptions and FTE allocations for the core team and computing support are presented in the main text (Section 7).

CORE DAPP TEAM

The core preservation team comprises full-time specialists responsible for building and operating the preservation infrastructure. Their roles and FTE allocations are detailed in Section 7.1.1 and accompanying tables.

ESSENTIAL FUNCTIONS WITHIN EXPERIMENTS

Each RHIC experiment plays a central role in ensuring long-term preservation. While coordination is supported through the central DAPP team, meaningful preservation requires contributors from within the collaborations to integrate preservation into ongoing analysis, publication, and documentation efforts. These activities are typically carried out as part of existing responsibilities (see Section 7.1.2).

Key Roles Within Experiments

The following functions are critical for implementing preservation at the experiment level and are **not detailed in the main text**. They reflect practical needs identified during planning discussions and represent areas where experiment-side engagement is essential:

- Scientific Knowledge Preservation: Identify and validate important datasets, code, and documentation with input from domain experts. Capture the reasoning behind key analyses and ensure that metadata is accurate, complete, and useful for future users.
- Code Preservation: Inventory legacy code and dependencies; build containers or other reproducible environments; document how analysis workflows are run and maintained.
- Engagement & Training: Develops training materials and encourages community participation in preservation practices. Promotes institutional memory and sustainable engagement.
- UX & Documentation Design: Supports interface design in collaboration with the Core Team; standardizes documentation; improves usability based on user feedback.
- Impact Analysis: Tracks usage, citations, and data reuse; compiles success stories and reports to stakeholders; identifies emerging community needs.

The main challenge is maintaining engagement and support for these roles as experiment activity declines. **These tasks require sustained effort and funding, they cannot be handled effectively on a volunteer basis alone.**

Computing Data Center Support

The Scientific Computing and Data Facility (SCDF) provides essential technical and operational support, including dedicated DAPP roles and generic computing services. Staffing details and transition planning considerations are outlined in Section 7.1.3 and Section 7.1.4.

Transition Planning

A formal transition plan will ensure the continuity of critical services and funding beyond Phase I, including the identification of infrastructure needs and the integration of specialized expertise into DAPP for Phase II.

Appendix D HARDWARE RESOURCE REQUIREMENTS

COMPUTING REQUIREMENTS

During Phase I, sufficient CPU resources are expected to remain available to support both data reprocessing and user analysis. However, effective coordination between the experiments will be essential to ensure effective resource sharing and scheduling. It is assumed that the maintenance and operation of the hardware at the SCDF will be done under RHIC operation until 2030. Except for dedicated services and AI components, the DAPP will require dedicated funding starting in 2030.

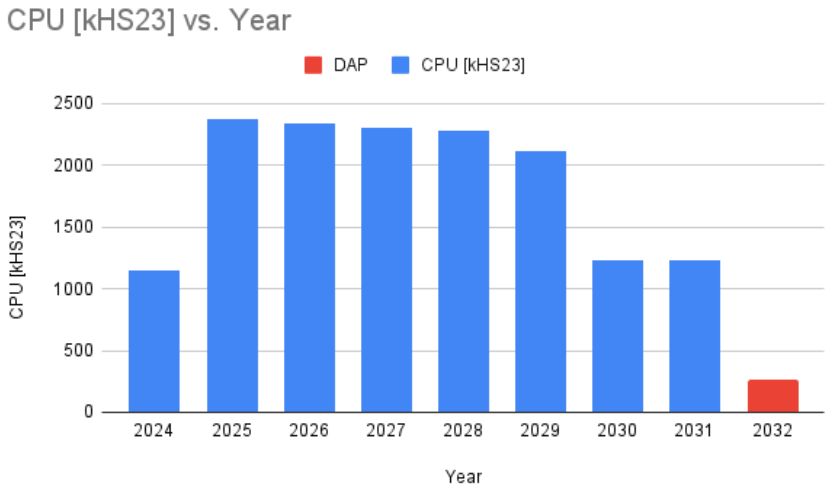
In Phase II, reprocessing will no longer be a significant activity. Instead, CPU demand will primarily be driven by data analysis. Based on historical trends, where analysis has accounted for up to 60% of STAR's total CPU consumption, we estimate that Phase II will require approximately twice the current level of STAR's typical analysis workload. This assumption will be refined as Phase II approaches. Opportunistic computing resources will supplement this core allocation from BNL's central facility or through ASCR facilities. Resource shortfalls are anticipated to begin around 2032, at which point additional CPU purchases will be necessary to sustain operations. Table 4 summarizes projected CPU needs over time.

Note: Hardware requirements for AI services are not included in this table. We plan to use AI platforms provided by the Computing and Data Science directorate and contribute to their maintenance and operation.

Table 4: Additional CPU requirement as a function of time

YEAR	2028	2029	2030	2031	2032
CPU [KHS06]	-	-	-	130	130

The timeline below illustrates the distribution of available CPU resources over time, with the hardware acquired through the DAPP initiative highlighted in red.



Disk Storage Requirements

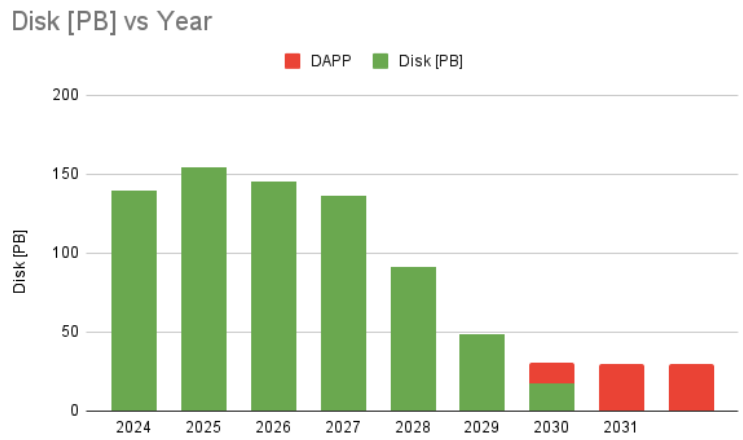
It is currently estimated that approximately 30 PB of disk storage will be required during Phase II to host the most used and scientifically valuable data objects. This estimate is preliminary; it will be refined based on the data volume of Run 2025, as well as ongoing efforts to optimize data structures through trimming and compression.

This estimate will be revisited regularly during Phase I as experience is gained. The details of disk storage requirements over time are indicated in Table 5.

Table 5: Additional Disk Storage requirement as a function of time

Year	2028	2029	2030	2031	2032
Disk [PB]	-	-	13	17	

The timeline below illustrates the distribution of available CPU resources over time, with the hardware acquired through the DAPP initiative highlighted in red.



TAPE STORAGE INFRASTRUCTURE

The current archive system uses multiple generations of LTO tapes. To ensure long-term accessibility and media compatibility, a full migration to LTO-11 is planned to begin around 2030, spanning approximately two years.

The migration will involve fully repackaging all archived tapes from earlier LTO generations over a two-year period. For the planned 750 PB of data, including 550 PB of raw data, current estimates indicate that approximately 45 LTO-11 drives will be needed for the raw data and 15 additional drives for the Analysis Objects. An equivalent number of older-generation drives will also be required to read the existing tapes. Additional drives may be needed for other data categories. Creating a second copy of any dataset will double the time needed to repackage that sample.

These same drives will also serve dual purposes by supporting data migration from tape to disk as users access preserved data on tape. Consequently, repacking activities may be delayed during periods of high demand for access.

To maintain one tape copy for each major data category (raw data, analysis objects, etc.), the total number of LTO-11 tapes to be purchased is summarized in TABLE 6, which is based on projected storage requirements and compression ratios.

Table 6: Requirements for the tape system as a function of time

YEAR	2028	2029	2030	2031	2032
LTO-11 Drives	-	-	60	-	-
LTO-11 MEDIA - RAW	-	-	4,700	4,700	
LTO-11 MEDIA – ANALYSIS OBJECTS			1,600	1,600	

TABLE 7 summarizes the hardware requirements for the RHIC DAPP.

Table 7: Hardware budget profile for RHIC DAPP

Component / [k$]	2026	2027	2028	2029	2030	2031	2032
Central Storage	$ -	$ -	$ -	$ -	$ 620	$ 800	$ -
CPU	$ -	$ -	$ -	$ -	$ -	$ 870	$ 870
Tape Drives	$ -	$ -	$ -	$ -	$ 300	$ -	$ -
Tapes Raw Data	$ -	$ -	$ -	$ -	$ 1,190	$ 1,190	$ -
Tapes Analysis Objects	$ -	$ -	$ -	$ -	$ 410	$ 410	$ -
Services and Miscellaneous	$ 300	$ 300	$ 300	$ 300	$ 300	$ 300	$ 300
Total	**$ 300**	**$ 300**	**$ 300**	**$ 300**	**$ 2,820**	**$ 3,570**	**$ 1,170**

Appendix E RISK MANAGEMENT ANALYSIS

The RHIC DAPP faces risks in four main areas that could impact the success of data preservation: data quality and processing, technology and infrastructure, institutional and funding, and user adoption and community. During Phase I, detailed risk registries will be developed for each category, including specific mitigation strategies, monitoring procedures, and contingency plans. Below is a summary of these risk categories along with key mitigation approaches.

GENERAL MITIGATION STRATEGIES

The RHIC DAPP faces several risk categories that could impact preservation success. Below is a summary of key risks with their corresponding mitigation strategies.

1. Hardware Failures

Risk: Storage hardware failures are expected over the preservation period. JBOD and ZFS are current technologies but can experience failures.
Mitigation: Regular maintenance, continuous monitoring, and relying on the experienced tape operations team. Preservation leverages existing SCDF infrastructure shared by other programs.

2. Data Corruption

Risk: Low-frequency data access increases the risk of undetected bit rot and file corruption, especially in large ROOT files.
Mitigation: Implement automated checksum verification on disks and comprehensive tape integrity checks during media migrations.

3. Disaster Events and Cybersecurity

Risk: Natural disasters, hardware failures, or cyberattacks could cause prolonged service disruptions or data loss

Mitigation: Deploy defense-in-depth security together with BNL's cyber team, incident response plans, and regular security audits. Leverage physically distributed storage, system snapshots, and automated backups. Develop and maintain a formal Disaster Recovery Plan. Test recovery procedures periodically to validate readiness.

4. Media Obsolescence

Risk: Tape formats may become unreadable as drives are phased out.
Mitigation: Plan regular migrations to current storage media based on annual Technology Watch reports tracking industry trends.

5. Software Dependencies

Risk: Legacy and hardware-specific software may become incompatible with future systems.
Mitigation: Preserve full software environments using container images, supported by detailed documentation and regular compatibility testing.

6. Computing Platform Changes

Risk: Hardware architecture shifts (e.g., from x86 to ARM) may affect software compatibility. **Mitigation:** Build container images targeting multiple architectures, conduct annual cross-platform testing, and use virtualization for legacy code fallback.

7. Funding Discontinuation

Risk: Loss of Phase II DOE funding threatens long-term preservation. **Mitigation**: Develop a scalable service model with defined preservation tiers to ensure core capabilities are maintained under constrained budgets. Prepare a public-access package as a baseline safeguard (like a website or InvenioRDM repository). Proactively engage with DOE by providing clear metrics on scientific impact and reuse to justify continued investment. Simultaneously, cultivate partnerships with universities and computing consortia to explore co-funding and fallback hosting options that strengthen long-term sustainability.

8. Personnel and Knowledge Loss

Risk: Staff turnover risks loss of institutional knowledge during phase transitions. **Mitigation:** Document critical procedures thoroughly, promote cross-training for operational resilience, and identify and work with key personnel to ensure continuity of knowledge and expertise.

9. Institutional Changes at BNL

Risk: Shifts in laboratory priorities or IT services may reduce support. Mitigation: Regular engagement with BNL management and ITD, supported by clear impact metrics, to demonstrate the value of sustained support. Build partnerships across departments and with external institutions and remain responsive to organizational changes. The modular design ensures the system remains adaptable if priorities shift.

10. Infrastructure Aging

Risk: Computing hardware and tape systems will become obsolete by 2030 without a refresh. **Mitigation:** Integrate preservation hardware needs into existing SCDF lifecycle management and refresh plans.

11. Low User Adoption and Knowledge Gaps

Risk: Limited future use and expertise may reduce preservation impact. **Mitigation:** Employ user-centered design, AI assistance, expert support, and educational partnerships to sustain user engagement and knowledge transfer.

12. Competition from EIC

Risk: EIC's upcoming data and community focus may reduce interest in RHIC data. **Mitigation**: Emphasize RHIC's unique and comprehensive physics coverage as a critical historical baseline that complements future EIC data. Underscore that RHIC will remain the principal data source through DAPP Phase II, ensuring continuity in key research domains until EIC becomes operational.

13. Experiment Differences

Risk: Diverse workflows across RHIC experiments complicate unified preservation. **Mitigation:** Provide common infrastructure for storage, security, and metadata while maintaining experiment-specific environments and documentation.

14. Governance and External Dependencies

Risk: Multiple stakeholders and reliance on external services pose coordination and continuity challenges.
Mitigation: Establish transparent decision-making processes and maintain backup plans, including local alternatives for critical external services.

15. AI System Reliability and Drift

Risk: AI-powered tools may become outdated, misaligned with user needs, or inaccurate as technologies and user expectations evolve.

Mitigation: Establish a technology evolution watch process for tracking advances in AI/ML infrastructure. Integrate automated workflows for updating AI training corpora with newly published data, metadata, and user interactions. Implement regular user feedback surveys and dedicated feedback channels to guide continuous improvement and maintain relevance.

16. Legal and Regulatory Changes

Risk: Future changes in data privacy, intellectual property laws, or international data-sharing regulations could impact long-term access or preservation workflows.

Mitigation: Conduct periodic legal and regulatory reviews (e.g., every 3–5 years) to assess alignment with evolving policies. Maintain institutional liaison with BNL's legal and compliance teams to ensure timely adaptation and compliance. Document the provenance and rights for preserved datasets to facilitate compliance with evolving requirements.

References

[1] U.S. Department of Energy, *DOE Order 241.1C: Scientific and Technical Information Management*, https://www.directives.doe.gov/directives-documents/200-series/0241.1-border-c

[2] Office of Science and Technology Policy (OSTP), *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research*, August 25, 2022 ("Nelson Memo"), https://www.whitehouse.gov/wp-content/uploads/2022/08/OSTP-Public-Access-Memo.pdf

[3] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-M-2, Magenta Book, June 2012.

[4] CoreTrustSeal Standards and Certification Board. (2023). CoreTrustSeal Trustworthy Data Repositories Requirements 2023–2025. https://doi.org/10.5281/zenodo.7051012

[5] Wilkinson et al., *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[6] CERN Open Data Portal, https://opendata.cern.ch

[7] HEPData Repository, https://www.hepdata.net

[8] Šimko, T., Nielsen, L. H., Marian, L., et al. (2020). InvenioRDM: A modern research data management repository platform. https://doi.org/10.5281/zenodo.3661596

[9] Šimko, T., Heinrich, L., Hirvonsalo, H., et al. (2019). REANA: A System for Reusable Research Data Analyses. EPJ Web of Conferences, 214, 06034. https://doi.org/10.1051/epjconf/201921406034

[10] Amstutz, P., Crusoe, M. R., Tijanić, N., et al. (2016). Common Workflow Language, v1.0. https://doi.org/10.6084/m9.figshare.3115156.v2

[11] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Wu, S., Xiong, Y., Zhang, J., Lin, K., & Zhou, M. (2025). *Retrieval-Augmented Generation for Natural Language Processing: A Survey*. arXiv preprint arXiv:2407.13193.

[12] Wikipedia contributors. (2025, June). *Model Context Protocol*. Retrieved from https://en.wikipedia.org/wiki/Model_Context_Protocol

[13] Basaglia, T., Bellis, M., Blomer, J. *et al.* Data preservation in high energy physics. *Eur. Phys. J. C* 83, 795 (2023). https://doi.org/10.1140/epjc/s10052-023-11885-1

[14] ICFA Panel on Data life cycle, https://icfa.hep.net/icfa-panel-on-the-data-lifecycle/

[15] Red Hat, Inc. (n.d.). *What is OpenShift?*. Retrieved June 25, 2025, from https://www.redhat.com/en/technologies/cloud-computing/openshift