August, 2025

DAPP RHIC Review Report

The DAPP Evaluation Committee (DEC)[1] : Cristinel Diaconu (Chair, CPPM, CNRS/IN2P3 and Aix-Marseille Université), Kati Lassila-Perini (Helsinki Institute of Physics, Finland), Achim Geiser (DESY, Hamburg), Simone Campana (CERN, Geneva), Ralf Seidl (RIKEN, Japan), Ulrich Schwickenrath (CERN, Geneva).

Executive Summary

The Relativistic Heavy Ion Collider (RHIC) physics program stands as a monumental achievement for humanity and a cornerstone of international leadership in scientific research. Its current accomplishments are already impressive, and continued investment and exploitation of this program, through a dedicated data preservation plan such as DAPP, are expected to significantly amplify research outcomes, beyond the active running phase of the experiments.

This approach has been validated over the past decade by similar initiatives at HERA, BaBar, and LEP, as documented within the DPHEP collaboration. These efforts have demonstrated that a focused, realistic, and well-targeted investment in data preservation yields substantial scientific returns at a low cost[2].

The Data and Analysis Preservation Plan (DAPP) is a robust and forward-thinking strategy. It fosters a unified and coherent understanding between experimental teams and the computing center. The plan is timely, proposed on one hand during a period of advanced maturity in physics analysis and on the other hand ahead of extensive data collection. This dual timing advantage allows for a solid basis for further organizational and technological developments, thereby enhancing the overall long-term robustness.

The DAPP showcases a strong commitment from the host laboratory structures, which is a highly positive aspect. Besides the scientific return, the laboratory will benefit by installing new practices compatible with open science and FAIR approaches.

The project is not only a preservation and re-exploitation initiative, but also includes innovative elements, such as an AI-based development for knowledge preservation. The investigation of AI techniques in the field of knowledge preservation will not only improve the long-term data analysis effectiveness, but it is also likely to have further applications for the ongoing and future projects in high-energy physics and beyond, in other science and cultural projects or industry. The plan is very well structured with a staged approach, building on existing frameworks and experiences. It is strongly linked to international collaborations, thereby gaining in visibility and ensuring long-term sustainability.

The plan benefits from a long-term vision at the host lab, with the Electron-Ion Collider flagship project to become operational in a decade and that overlaps in physics interests. This common interest encourages continued collaboration and leverages knowledge preservation at the host lab. DAPP will certainly benefit EIC not only through common scientific interest, training generation of students and ensuring a data-based science production through the “dark times” with no collisions at BNL, but also by installing better practices and innovative approaches in data collection and analysis at the host lab.

In summary, the DAPP project is not only excellent and innovative, with a strong potential to produce cost-effective cutting-edge science based on RHIC data, but also positions the host laboratory BNL and the United States as leaders in this field. The committee supports very strongly this project and formulates a number of comments towards maximizing its impact.

[1] Process: Charge received June 10, 2025; Project documents available June 19; DEC Preparatory Meeting June 25; Questions by the DEC submitted on June 27; DAPP Review Meeting July 1st, 2025.

[2] DPHEP Collaboration • T. Basaglia (CERN) et al., Eur.Phys.J.C 83 (2023) 9, 795.

Introduction and general remarks

This evaluation report assesses a proposal centered on the continued exploitation, preservation, and evolution of data and infrastructure from RHIC collider at BNL. The unique and extensive datasets generated over more than two decades of operations - through major experiments such as PHENIX (2000–2016), STAR (2000–2025), sPHENIX (2023–2025), BRAHMS (2000–2006), and PHOBOS (2000–2005) - represent an invaluable scientific asset. Notably, these datasets continue to yield high-impact publications well beyond the end of active data taking, underlining « in situ » their long-term scientific relevance.

The upcoming Au+Au run is expected to produce nearly half of the total available data, further reinforcing the importance of long-term data preservation and access strategies. In parallel, the transition toward the Electron-Ion Collider (EIC), with construction planned to begin in 2026, marks a significant shift in the research landscape. The proposed initiative makes a compelling case for bridging this transition period, through sustained access to legacy data and the expansion of the physics case, including potential synergies with future EIC developments.

While the scientific ambition is clear, the evaluation highlights several critical aspects that will require careful attention. The technical foundations for software and source code preservation, and the maintaining the ability to rebuild the software stack from these sources, appear underdeveloped and must be more robustly addressed. Risk management strategies—especially in light of potential funding fluctuations and the need for prioritization—should be further elaborated. Clear engagement and alignment with the experimental collaborations are essential for the success of the initiative.

Artificial Intelligence (AI) is proposed as a key enabler for enhancing data accessibility and analysis. The committee appreciates the innovative and forthcoming aspects of using this novel technologies for developing a versatile, complete and efficient data preservation system. The mix of AI-based approaches and solutions is sound. However, the evaluators underscore that AI cannot substitute the indispensable, manual work of metadata curation and quality control. AI tools may assist and validate, but curated, well-documented data and usage instructions remain a foundational prerequisite.

The proposal's ability to keep the project at the forefront of technological innovation will depend critically on the willingness and capacity of experimental teams to engage in this demanding process. The extra-risk taking for the long-term system stability should be carefully evaluated while relying on AI developments. The existing STAR AI-based prototype is a very good asset, however the difficulty of a transition to a long term model is not to be underestimated.

Phase I of the project will need to demonstrate the feasibility and effectiveness of all core components, with a readiness to refine or remove underperforming elements in Phase II. Overall, the project shows promise for sustaining scientific output, fostering technological advancement, and enabling continuity during a pivotal transition in the field of nuclear and particle physics.

This report is organized following the five questions formulated in the mandate.

1. Has the DAPP effectively identified and plan to preserve the most valuable scientific assets and legacy from the RHIC experiments?

Findings

The proposal presents a strong and well-structured plan, clearly organized into two distinct phases. This staging allows both for innovation in the first phase and for stabilization and possible planning adjustments for the second (long term) phase.

The datasets produced by the RHIC experiments are unique and not reproducible in the medium to long term. Their long-term value for the broader scientific community is evident and the host laboratory is encouraged to pursue their long-term preservation.

Continued exploitation of these datasets is not only justified but strongly encouraged, as they remain open to novel analyses and future scientific opportunities.

However, the current proposal would be reinforced by a more detailed physics case. The scientific motivation, while implicitly strong, should be made more explicit to strengthen the project's coherence and impact. Past experience shows that steady consideration of physics goals and careful planning of physics analysis allows better identification of resources and stimulates further physics developments beyond initial plans.

The structure of the RHIC Data Archiving and Preservation Project (DAPP) appears to be well-suited for achieving the goal of long-term data preservation. The governance model of the RHIC DAPP is designed to facilitate effective decision-making, ensure accountability, and incorporate best practices and expert advice, all of which are crucial for the success of a long-term data preservation project.

Moreover, the proposal correctly identifies the crucial aspect of the institutional support, for which adequate instances are in place at the host laboratory. A special attention has to be given to the interactions with users, in particular with the collaborations that will most likely conserve the scientific supervision of the data reuse in the medium and longer terms.

Comments

The proposal should more clearly and forcefully present the scientific opportunities—the "great physics case"—that justify the effort and investment. This includes demonstrating how preserved data can support innovative or high-priority physics questions.

The roles and responsibilities of the collaborating experiments and institutions should be more precisely defined to ensure accountability and effective implementation across both phases.

The Electron-Ion Collider (EIC) should be explicitly considered as a future stakeholder and potential user of the preserved data and tools. The proposal would benefit from identifying how it can serve EIC-specific physics goals and build early connections with the EIC community to maximize relevance and uptake.

2. Will the proposed infrastructure enable both verification of published results and new analyses by external researchers?

Findings

The proposed data workflow is well-conceived and, in principle, supports full reproducibility of scientific results.

It remains unclear how feedback from internal researchers has been integrated into the development of the workflow and whether their practical needs and insights are being sufficiently addressed.

Questions persist regarding the applicability of this workflow to all previously published papers, particularly in light of legacy systems and evolving technical requirements.

The proposal rightly positions the project and its collaborators at the technological forefront through the integration of AI, with potential empowering of the published results verification and new analyses by external researchers – the profile and definition of which has to be clarified (see below).

Comments

The team should plan to rely as much as possible on the common infrastructure already available at the host laboratory, minimizing dependence on hardware- or software-specific solutions that may limit scalability or sustainability in the long term.

Long-term support structures and services should be clearly outlined, ensuring the workflow remains usable, maintainable, and accessible well beyond the initial project phases.

Greater clarity is needed on how the workflow will be applied retroactively to existing publications, and whether it meets the technical and practical requirements for reproducibility across the entire legacy dataset.

The indispensable work of curating and reviewing data must be carried out by the experiments themselves and constitute a key element of the whole plan. While AI tools can assist and validate certain processes, they cannot replace the essential human expertise required for data quality assurance and metadata completeness.

The notion of “external user” needs to be carefully defined, as a function of various use case and connections with the collaborations. In that sense, the definition of the use case for new users need some structural approach within the collaborations themselves in collaboration with the DAPP core team. The use case for an independent usage of data requires more specific connections with an Open Science approach.

Therefore, the collaborations should propose and implement a workflow for new collaborators and groups joining the collaborations all through the long-term preservation phase, and prepare the necessary coaching, training and scientific supervision for successful data analysis. The use of AI looks promising from that point of view. It should be complemented by an explicit evolution of the collaborations organization for the long term.

3. Are proposed data curation practices sufficient to ensure long-term usability and discoverability of RHIC data?

Findings

The proposal presents a comprehensive and well-developed plan that addresses multiple levels of usability, from infrastructure to individual user/analysis components.

It thoughtfully incorporates Artificial Intelligence (AI) as a supporting technology to enhance data accessibility and analysis.

FAIR (Findable, Accessible, Interoperable, Reusable) data principles are strongly embedded in both the strategic vision and the technical infrastructure, demonstrating a clear alignment with best practices in data stewardship.

The participating experiments are already engaged in efforts to preserve analysis workflows and support reproducibility, indicating a strong foundation for the proposed developments.

Comments

Priority should be given to the co-development and implementation of reproducibility workflows in close coordination with the experiments, ensuring that these are grounded in actual use cases and operational practices.The preservation of simulated data is to be considered an important part of the project as well.

(Meta)data curation must begin as early as possible within the experiments themselves. This is a critical prerequisite for the success of the project and cannot be outsourced or delayed, as well-structured metadata is essential for both reproducibility and effective AI integration.

4. Are the proposed FTE allocations and infrastructure requirements realistic for both the initial and sustained implementation phases?

Findings

The proposed Full-Time Equivalent (FTE) allocation appears reasonable and well-aligned with the scope of the planned activities. The committee took note of the global commitment of the collaborations and considers that this is already a decisive step.

However, the contribution from the experimental collaborations is currently expressed only in terms of general commitments, without concrete FTE allocations. Mutual expectations between the project team and the collaborations require clearer definition to ensure an effective collaboration.

It remains to be clarified to what extent the collaborations' contributions will be critical to the long-term success of the project, and whether this dependency is adequately managed.

Flexibility in task allocation should be preserved to allow for some degree of effort transferability among participants.

Comments

It is essential to ensure that the human resources dedicated to the project are adequately complemented by measurable and active contributions from the collaborations.

At minimum, the expectations from each participating collaboration should be formalized—defining roles, responsibilities, and anticipated effort levels—to ensure shared understanding and sustained engagement throughout the project lifecycle.

5. Has the plan identified risks and outlined suitable mitigation strategies?

Findings

At this stage of preparation, the proposal provides a solid foundation and has addressed a vast amount of aspects. The project includes many positive aspects that will make the long-term preservation robust, such as the final, comprehensive reprocessing of sPHENIX and STAR data to produce consistent and well-documented datasets.

The initial planning is well-structured and reflects a thoughtful approach. In the area of risk assessment and mitigation there are a few items that could still be addressed, either upfront or during the first phase of the project.

Comments

The risk analysis should be further developed and refined to account for technical, organizational, and funding-related uncertainties, especially given the project's multi-phase structure and dependence on collaboration inputs.

A task sharing within and across the collaborations and the Computing and Data Science Directorate, at both technical and more physics related levels, can be an interesting path to follow in order to mitigate the main risks related to knowledge dissipation in the long term.

Conclusions

Preserving RHIC data will produce new science results and will maximize the scientific outcome of the massive investment made for this unique experimental program.

This is a strong and forward-looking data preservation project that goes well beyond safeguarding past scientific output. It leverages a unique and non-reproducible body of experimental data, accumulated over decades, and proposes an ambitious yet well-grounded plan to ensure its continued accessibility and scientific impact.

Importantly, the project does not limit itself to traditional preservation goals—it also aims to innovate, notably through the integration of Artificial Intelligence tools to enhance data usability, reproducibility, and future discovery potential.

The proposal is strategically structured, with a phased approach that balances ambition with feasibility. It demonstrates alignment with FAIR data principles and reflects a thoughtful consideration of emerging community standards. The inclusion of AI places the project at the forefront of technological developments in data-intensive science.

A number of comments have been provided, mostly oriented toward refinement and optimization. These include clarifying the contributions and expectations of collaborating experiments, prioritizing early metadata curation efforts, expanding the reproducibility framework, and improving the risk analysis. These are not fundamental weaknesses, but areas where further elaboration and rigor would strengthen the overall proposal and reduce long-term risks.

The evaluators commend the team for the depth and vision of their approach and encourage continued leadership and engagement in the global data preservation landscape.

Recommendation

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