RHIC Data and Analysis Preservation Round Table

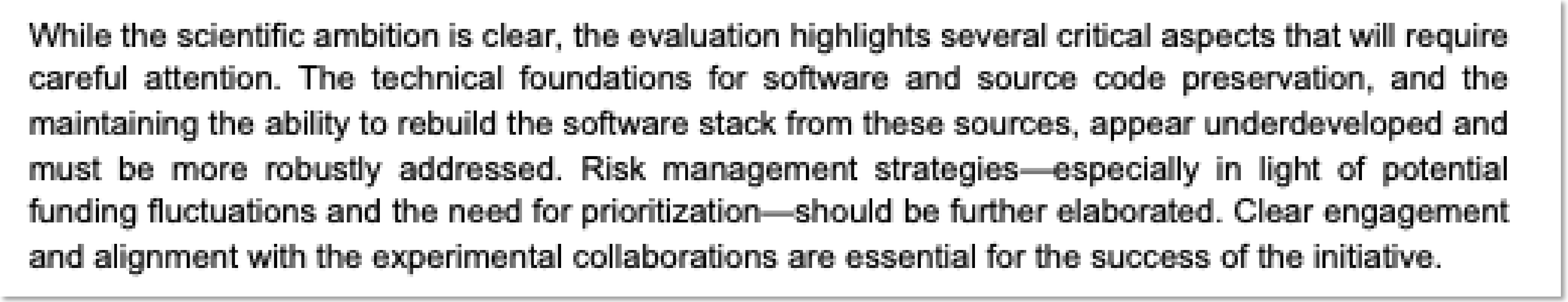
09/25/25

Introduction – Follow up on review report

Plan for today

- Software & source code
- Funding risks & prioritization
- Engagement with collaborations
- Integration into workflows & publications
- Metadata challenges & solutions

Follow up on DAPP review



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.

Three priorities have been identified

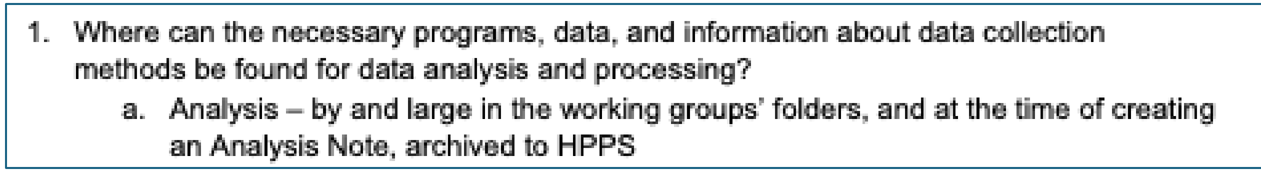
1. Software and source code preservation, ability to rebuild the software from sources
2. Risk management: funding fluctuations and prioritization
3. Engagement and alignment with collaborations

1- On software

- *“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.”*
- Some information about software and source code is available in the answers to the early 2025 questionnaires
- Some thoughts were presented by Jerome in April about code repository
- Current software preservation foundations are underdeveloped.

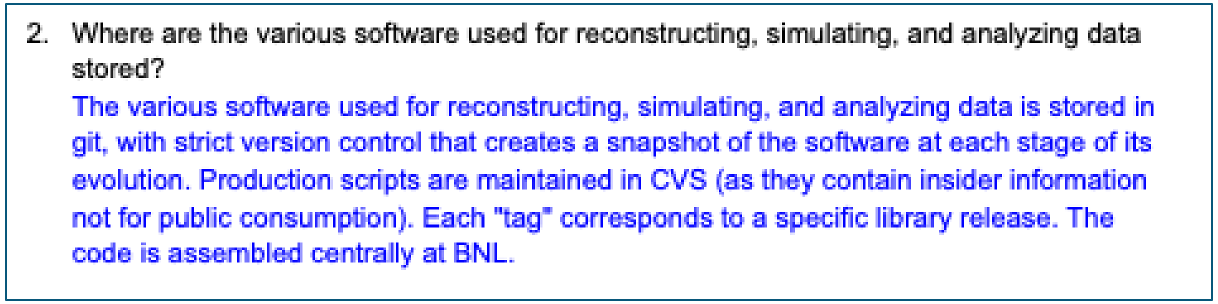
Code repository

- PHENIX answers
  - Analysis software is archived in HPSS
  - Is it available on a Git repository?



1. Where can the necessary programs, data, and information about data collection methods be found for data analysis and processing?
   a. Analysis – by and large in the working groups' folders, and at the time of creating an Analysis Note, archived to HPPS

- STAR answers
  - Software in https://github.com/star-bnl
  - All analysis software ?
  - How is it built ?



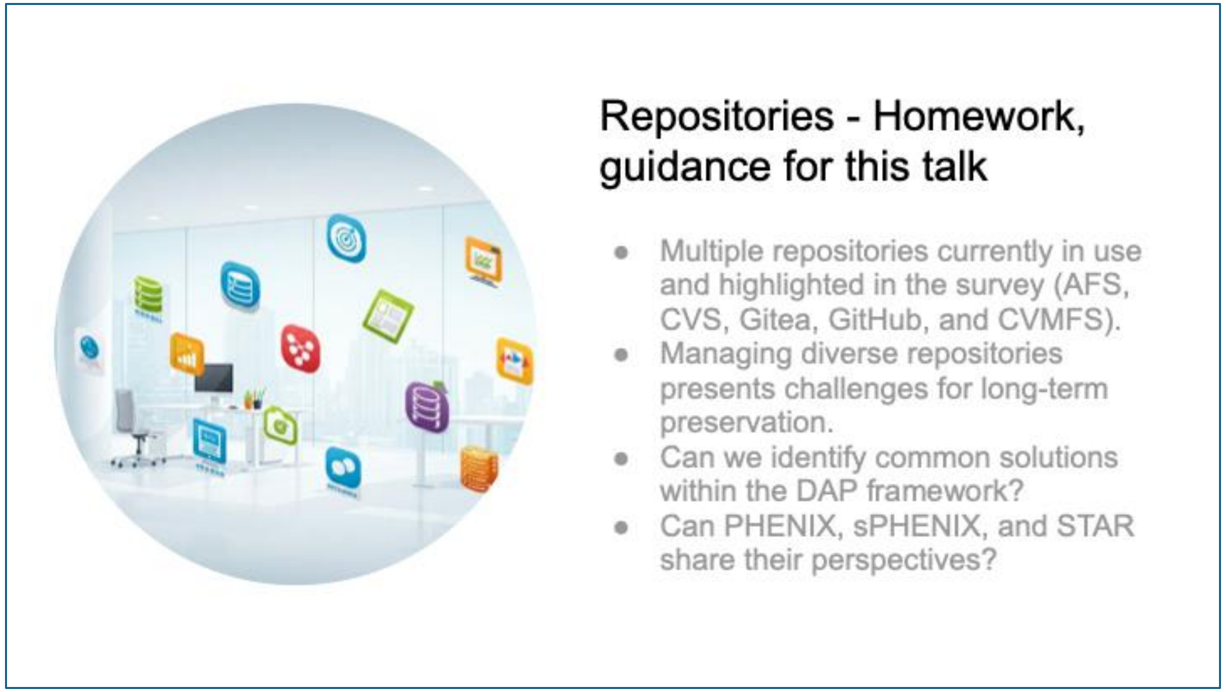
2. Where are the various software used for reconstructing, simulating, and analyzing data stored?
   The various software used for reconstructing, simulating, and analyzing data is stored in git, with strict version control that creates a snapshot of the software at each stage of its evolution. Production scripts are maintained in CVS (as they contain insider information not for public consumption). Each "tag" corresponds to a specific library release. The code is assembled centrally at BNL.

- sPHENIX
  - Software in https://github.com/sPHENIX-Collaboration
  - same questions as for STAR

Needs: reproducible builds, sustainable access, documentation

Jerome 4/24/25
https://indico.bnl.gov/event/26753/

Jerome’s talk



Repositories - Homework, guidance for this talk

- Multiple repositories currently in use and highlighted in the survey (AFS, CVS, Gitea, GitHub, and CVMFS).
- Managing diverse repositories presents challenges for long-term preservation.
- Can we identify common solutions within the DAP framework?
- Can PHENIX, sPHENIX, and STAR share their perspectives?

Due to the review in early July, this topic has not been actively followed up.

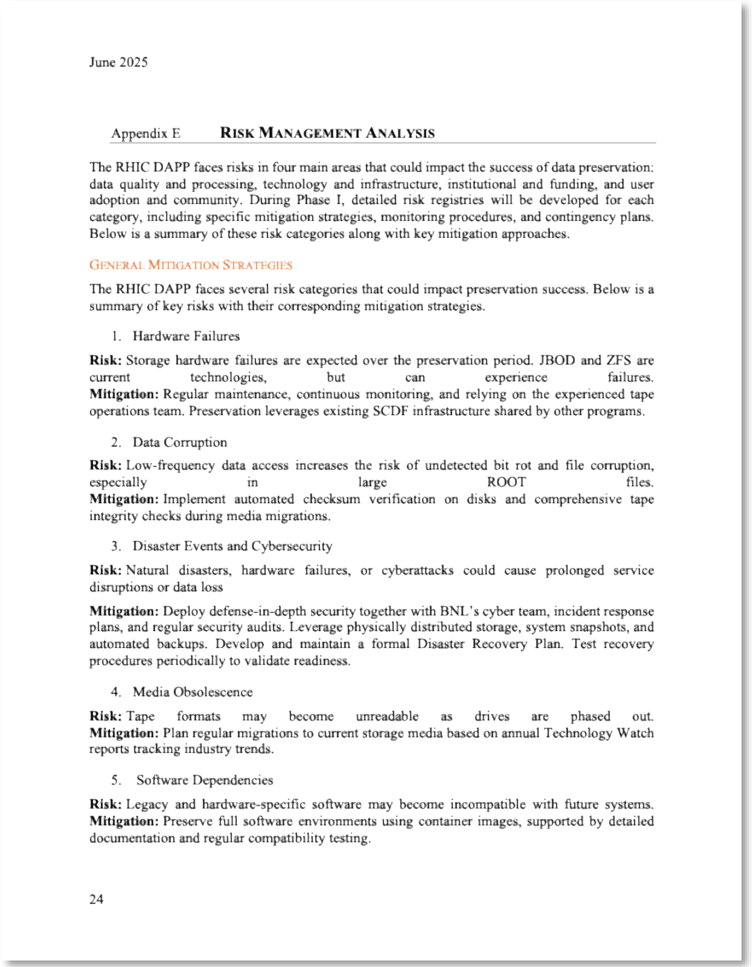
Software – proposed next steps

- Organize a dedicated topical meeting with all experiments
  - Goal: How to set up a reproducible software infrastructure for DAP?
- Gather and consolidate inputs on current practices, challenges, and requirements for software and source code preservation.
  - Comparison matrix mapping each experiment's: repositories, build systems, testing frameworks, software documentation, and preservation
  - Identify overlapping tools and divergent approaches
- Develop a plan for robust software preservation and reproducible builds.
  - Define minimum viable common solutions for DAP while allowing experiment-specific customization

2- Risk management

- *“Risk management strategies—especially in light of potential funding fluctuations and the need for prioritization—should be further elaborated”*
- The DAPP includes 16 detailed risks in appendix.
  - One risk addresses funding discontinuation during Phase II (long-term operation).
  - However, funding fluctuations during either phases and the need to highlight priorities are not directly addressed.
- Review highlighted missing elements:
  - Short-term funding fluctuations
  - Clear priority-setting framework
- Other major risks (missing DAPP):
  - Departure of key personnel
  - Inadequate funding

Appendix E



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

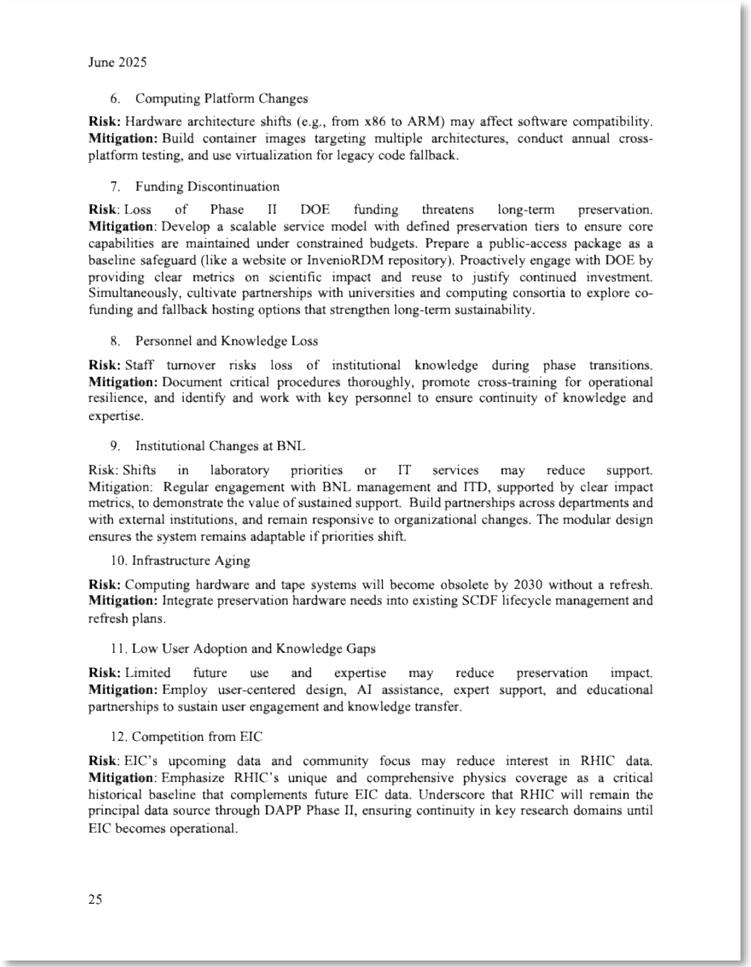
Mitigation: Deploy defense-in-depth security together with BNL's cyber team, incident response plans, and regular security audits. Leverage physically distributed storage, system snapshots, and automated backups. Develop and maintain a formal Disaster Recovery Plan. Test recovery procedures periodically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

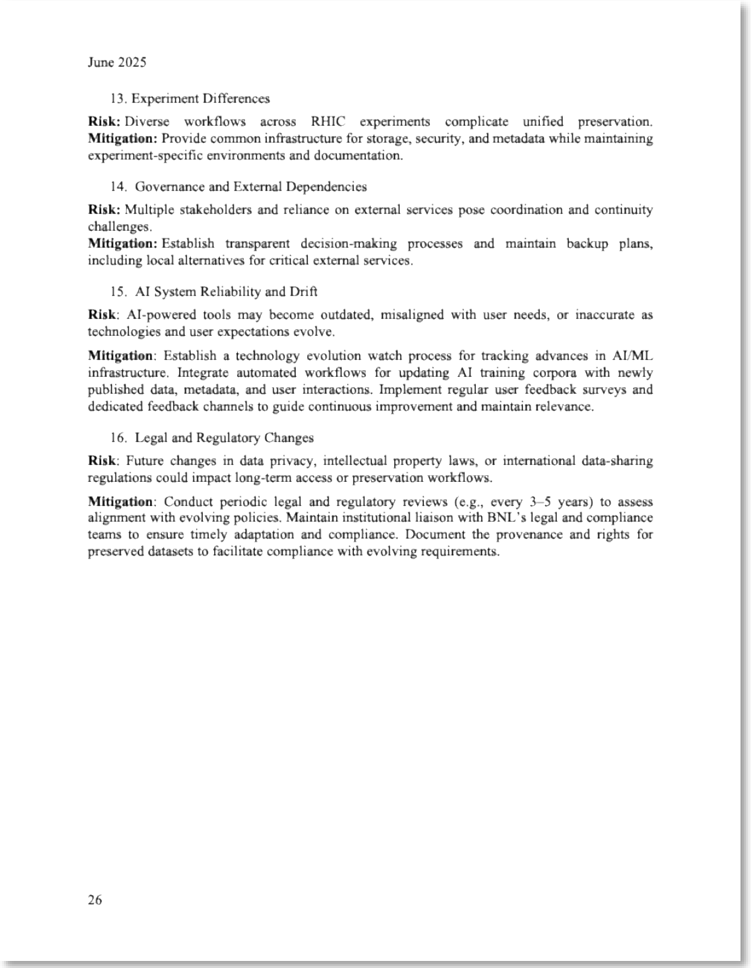
Risk: Computing hardware and tape systems will become obsolete by 2030 without a refresh. **Mitigation:** Integrate preservation hardware needs into existing SCDF lifecycle management and refresh plans.

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.

Follow-Up on Risk Management

- Jerome has proposed to follow up on this topic; an update will be presented at the next meeting
- Assess vulnerability to funding fluctuations
- Establish prioritization for activities
- Schedule a dedicated meeting for risk analysis and mitigation strategies.

3- Engagement from collaborations

- *“Clear engagement and alignment with the experimental collaborations are essential for the success of the initiative.”*
- Strong engagement and alignment with collaborations are essential.
- DAPP assigns 0.3 FTE/experiment for DAP coordination.
- Institutional commitment required to ensure sustainability

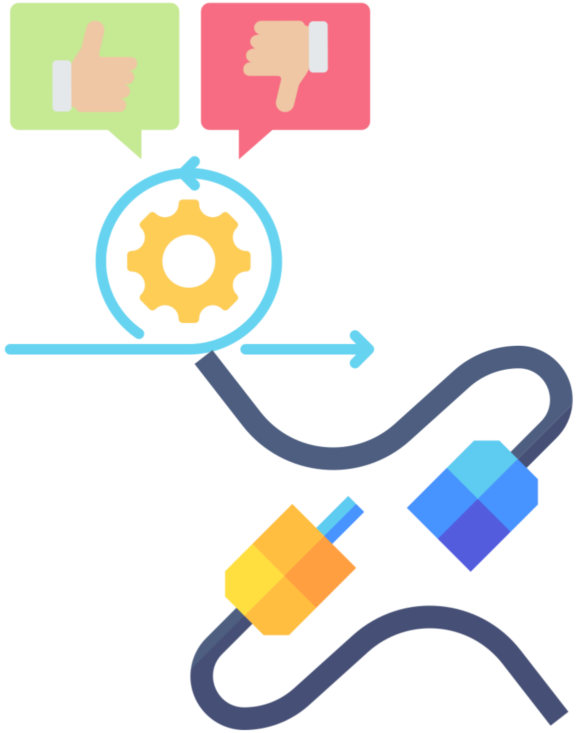
Engagement – proposed next steps

- Work with collaborations and BNL to identify DAP coordinators
- Make DAP part of everyday analysis workflows, not an add-on
- Embed preservation into publication processes

DAP integration

Analysis

Review process



Publication

DAP

Integrating DAP into the Publication Process

- Why?
  - Ensure reproducibility of published results
- How?
  - Preserve code, workflows, and data provenance with each paper
  - Use containers & documentation for reproducibility
  - Start with pilot publications
- Benefits
  - Seamless integration with existing publication flow
  - Long-term reproducibility guaranteed

What DAP implementation entails

- **Code & Software Preservation**
  - Version-controlled code and scripts
  - Containerized environments
  - Documentation of dependencies and configurations
- **Workflow & Data Preservation**
  - Step-by-step analysis procedures
  - Preservation of intermediate datasets (when needed)
  - Provenance tracking from raw data to results
- **Documentation & Metadata**
  - Detailed description of analysis
  - Software installation instructions
  - Parameter settings and calibration details
- **Integration with Publications**
  - Link preserved analysis to the corresponding papers

Objective: Guarantee long-term reproducibility and verification of results through systematic preservation of analysis workflows.

The metadata challenges

- We need to organize and make accessible scattered information in a clear and coherent manner for future users.
- Complexity: Multi-layered links (runs → datasets → analyses → publications)
- Variety: Raw data, ROOT files, calibrations, workflows, documentation, software versions
- Fragmentation: Custom, non-interoperable solutions across experiments
- Knowledge Loss: Expertise dispersing as operations end.

Metadata what we need

- Unified metadata system across experiments
- Simple, accessible to newcomers
- With enough information to perform re-analysis
- Standards-based for long-term viability
- Low maintenance post RHIC operation
- Open-source solutions exist: Apache Atlas, DataHub, OpenMetadata

Vincent agreed to follow up on this topic at a forthcoming meeting

Today

1. Update on DAP ChatBot – Alexandr Prozorov

- Next meeting: Thursday 10/01 – 10:00 AM EST