Brookhaven
National Laboratory

U.S. DEPARTMENT OF
ENERGY

Data and Analysis Preservation at RHIC

Eric Lancon (BNL)

PAC – Nov. 7, 2024

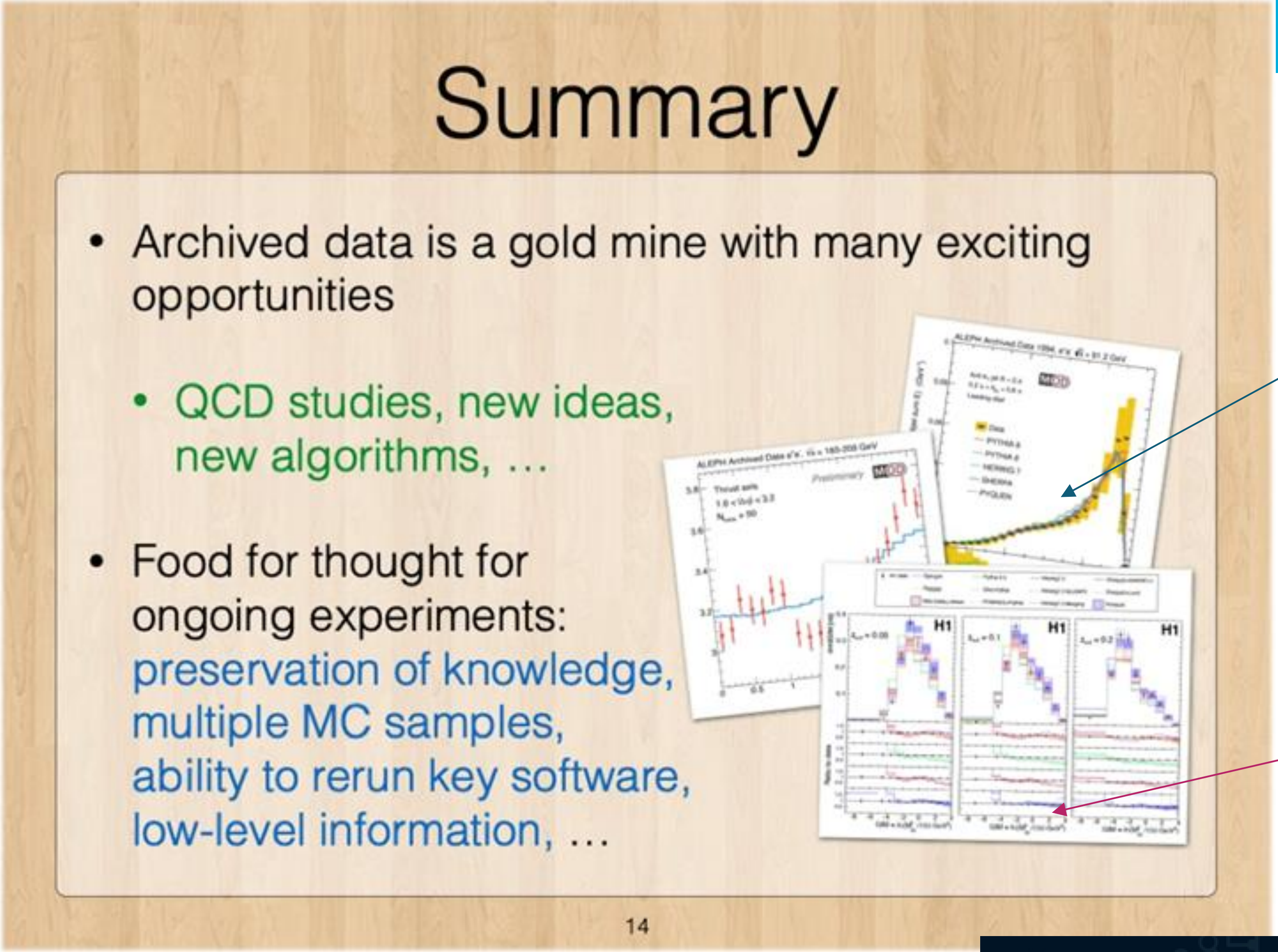
@BrookhavenLab

Importance of Data and Analysis Preservation

- The two decades of operation of RHIC have produced PetaBytes of invaluable data.
- Preserving and ensuring accessibility of these data and analyses is crucial to advancing scientific knowledge and fostering collaborative innovation.
- Data and Analysis Preservation:
  - Ensures long-term scientific impact, through potential future discoveries, verification, and reproducibility of results.
  - Supports the education and training of future scientists.
  - Enables reanalysis with new insights, ensuring data remains accessible as new theoretical advancements and discoveries emerge.

This effort aligns with DOE policy.

Old Data is a gold mine!



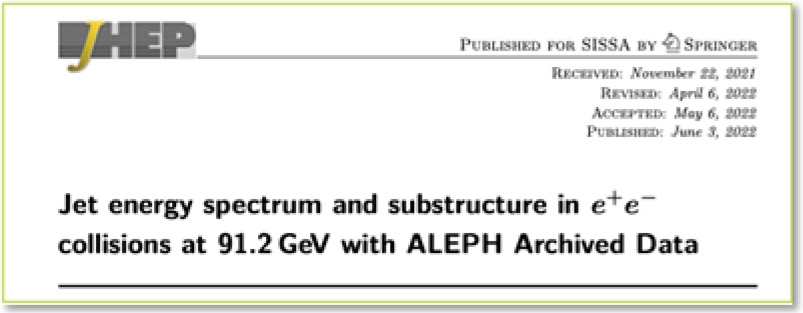
Summary

- Archived data is a gold mine with many exciting opportunities
  - QCD studies, new ideas, new algorithms, …
- Food for thought for ongoing experiments: preservation of knowledge, multiple MC samples, ability to rerun key software, low-level information, …

14

The data from the ALEPH and H1 experiments were reanalyzed using algorithms developed long after the data collection had concluded.

ALEPH experiment at LEP ended data taking in 2000
Analysis published in 2022

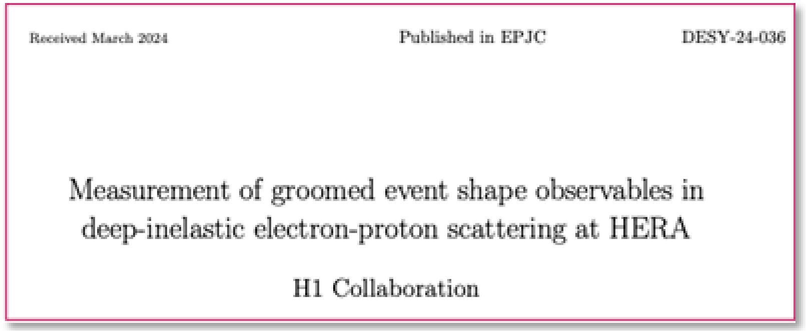


JHEP PUBLISHED FOR SISSA BY SPRINGER

RECEIVED: *November 22, 2021*
REVISED: *April 6, 2022*
ACCEPTED: *May 6, 2022*
PUBLISHED: *June 3, 2022*

Jet energy spectrum and substructure in e^+e^- collisions at 91.2 GeV with ALEPH Archived Data

H1 experiment at DESY ended data taking in 2007
Analysis published in 2024



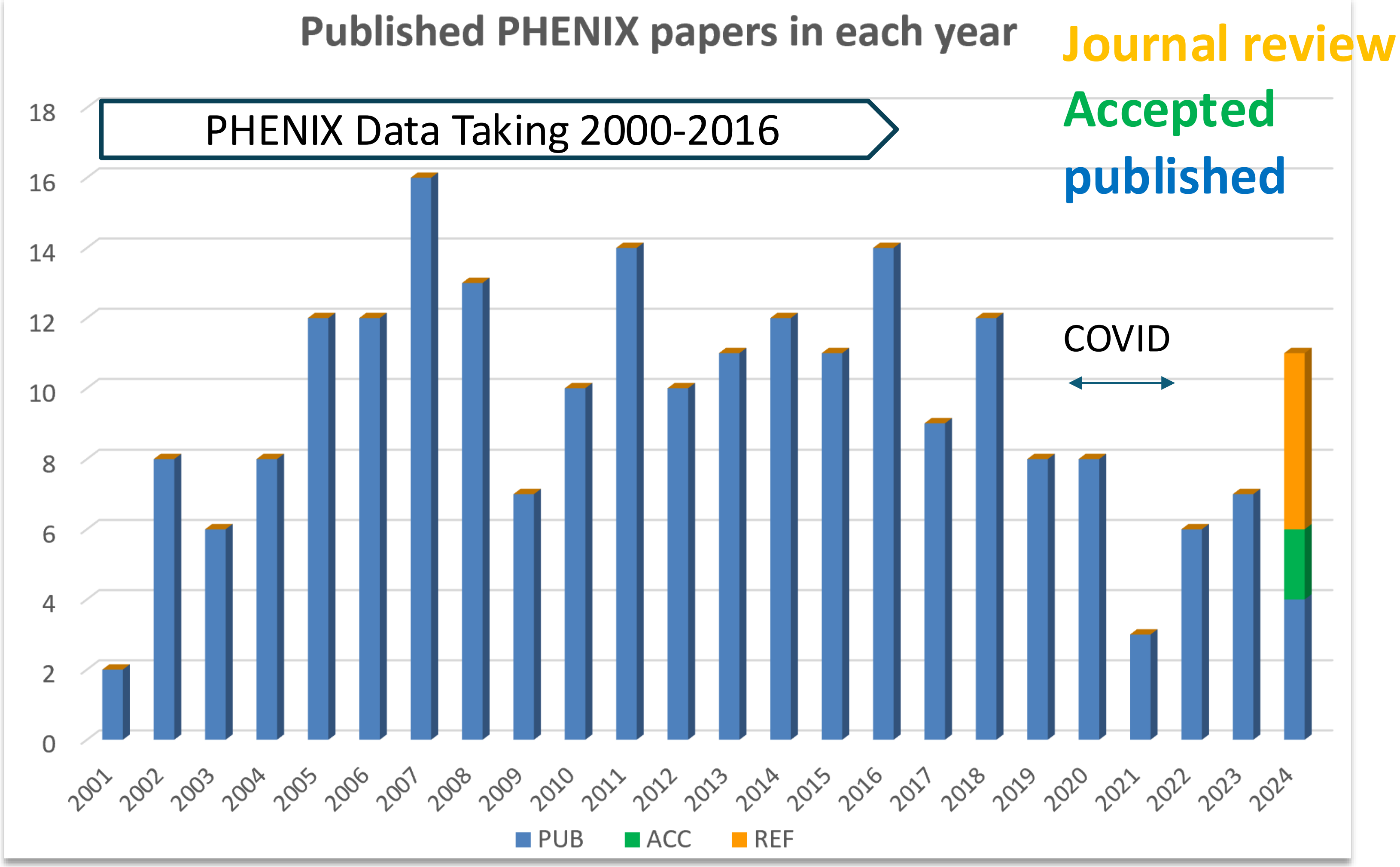
Received March 2024 Published in EPJC DESY-24-036

Measurement of groomed event shape observables in deep-inelastic electron-proton scattering at HERA

H1 Collaboration

Yi Chen (Vanderbilt), Peter Jacobs (LBNL) 2024 Jun 21, SANPC Workshop

Experiment publications extend well after the end of the data-taking: PHENIX example



Published PHENIX papers in each year

PHENIX Data Taking 2000-2016

Journal review

Accepted

published

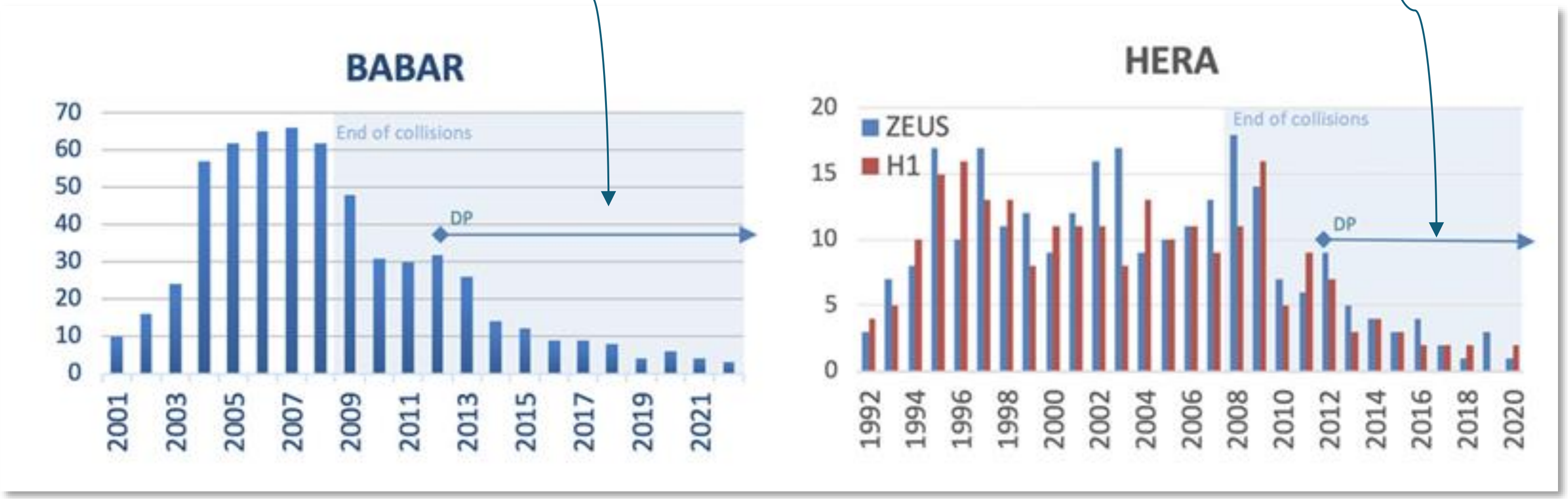
COVID

Experiment publications extend well after the end of the data-taking

EIC momentum is expected to drive increased re-analysis of HERA data.

33% of publications

15% of publications

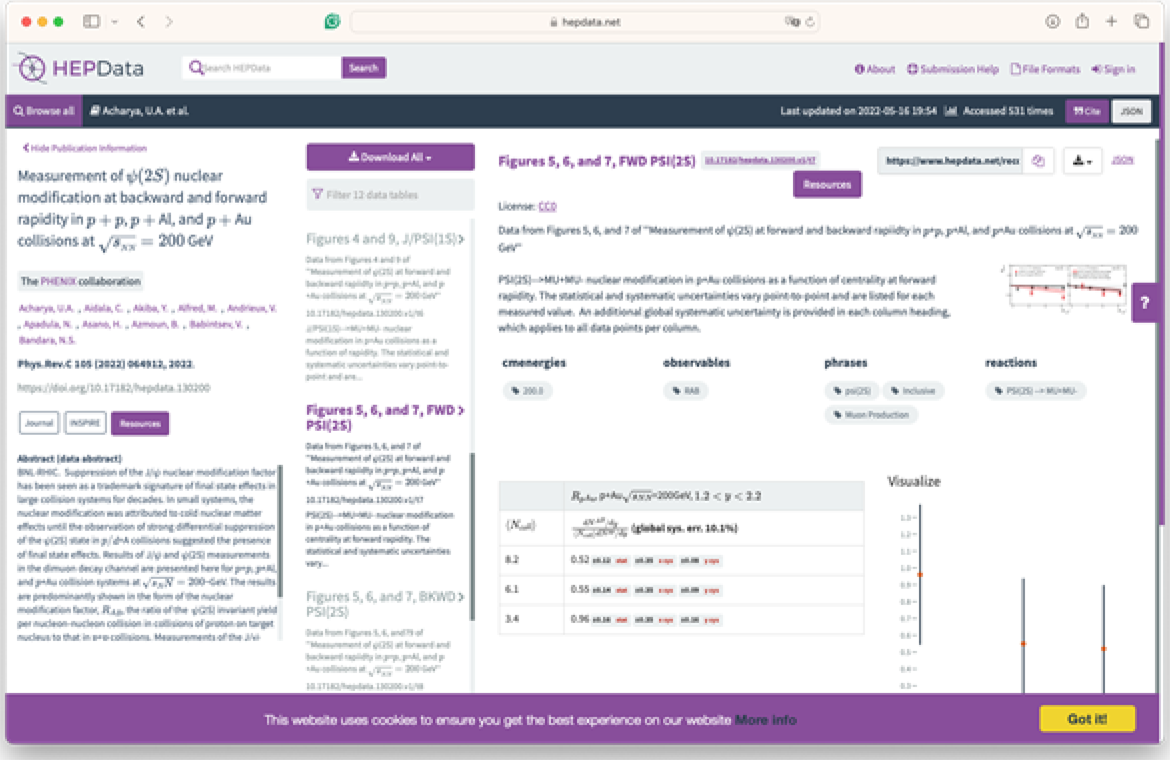


Several years after the end of data taking, an experiment enters the Data and Analysis Preservation mode.

Publications in Preservation mode may account for 10-20% of the total experiment publication set.

Brookhaven National Laboratory

Publishing is not Preserving Analyses



Last updated on 2022-05-16 19:54 | Accessed 531 times

Measurement of $\psi(2S)$ nuclear modification at backward and forward rapidity in $p+p$, $p+$Al, and $p+$Au collisions at $\sqrt{s_{NN}} = 200$ GeV

The PHENIX collaboration

Acharya, U.A. , Aidala, C. , Akiba, Y. , Alfred, M. , Andrieux, V. , Apadula, N. , Asano, H. , Azmoun, B. , Babintsev, V. , Bandara, N.S.

Phys.Rev.C 105 (2022) 064912, 2022.

https://doi.org/10.17182/hepdata.130200

Abstract (data abstract)
BNL-RHIC. Suppression of the J/ψ nuclear modification factor has been seen as a trademark signature of final state effects in large collision systems for decades. In small systems, the nuclear modification was attributed to cold nuclear matter effects until the observation of strong differential suppression of the $\psi(2S)$ state in p/d+A collisions suggested the presence of final state effects. Results of J/ψ and $\psi(2S)$ measurements in the dimuon decay channel are presented here for p+p, p+Al, and p+Au collision systems at $\sqrt{s_{NN}} = 200$ GeV. The results are predominantly shown in the form of the nuclear modification factor, R_{pA}, the ratio of the $\psi(2S)$ invariant yield per nucleon-nucleon collision in collisions of proton-on target nucleus to that in p+p collisions. Measurements of the J/ψ

Figures 4 and 9, J/PSI(1S)
Data from Figures 4 and 9 of "Measurement of ψ(2S) at forward and backward rapidity in p+p, p+Al, and p+Au collisions at $\sqrt{s_{NN}} = 200$ GeV"
10.17182/hepdata.130200.v1/t6
J/PSI(1S)-->MU+MU- nuclear modification in p+Au collisions as a function of rapidity. The statistical and systematic uncertainties vary point-to-point and are...

Figures 5, 6, and 7, FWD PSI(2S)
Data from Figures 5, 6, and 7 of "Measurement of ψ(2S) at forward and backward rapidity in p+p, p+Al, and p+Au collisions at $\sqrt{s_{NN}} = 200$ GeV"
10.17182/hepdata.130200.v1/t7
PSI(2S)-->MU+MU- nuclear modification in p+Au collisions as a function of centrality at forward rapidity. The statistical and systematic uncertainties vary...

Figures 5, 6, and 7, BKWD PSI(2S)
Data from Figures 5, 6, and7 of "Measurement of ψ(2S) at forward and backward rapidity in p+p, p+Al, and p+Au collisions at $\sqrt{s_{NN}} = 200$ GeV"
10.17182/hepdata.130200.v1/t8

Figures 5, 6, and 7, FWD PSI(2S) 10.17182/hepdata.130200.v1/t7

https://www.hepdata.net/rec

Resources

License: CC0

Data from Figures 5, 6, and 7 of "Measurement of ψ(2S) at forward and backward rapidity in p+p, p+Al, and p+Au collisions at $\sqrt{s_{NN}} = 200$ GeV"

PSI(2S)-->MU+MU- nuclear modification in p+Au collisions as a function of centrality at forward rapidity. The statistical and systematic uncertainties vary point-to-point and are listed for each measured value. An additional global systematic uncertainty is provided in each column heading, which applies to all data points per column.

cmenergies	observables	phrases	reactions
200.0	RAB	psi(2S) Inclusive Muon Production	PSI(2S) --> MU+MU-

	R_{pAu} p+Au $\sqrt{s_{NN}}$=200GeV, 1.2 < y < 2.2
$\langle N_{coll} \rangle$	$\frac{dN^{pA}/dy}{\langle N_{coll} \rangle dN^{pp}/dy}$ (global sys. err. 10.1%)
8.2	0.52 ±0.12 stat ±0.35 sys ±0.08 sys
6.1	0.55 ±0.14 stat ±0.35 sys ±0.08 sys
3.4	0.96 ±0.16 stat ±0.35 sys ±0.16 sys

This website uses cookies to ensure you get the best experience on our website More info

- **PHENIX and STAR** publish their figures and tables in HEPData, along with differential distributions of data and predictions. This approach provides the main inputs for statistical interpretation.
- **Preserving analyses** requires more than just publishing data.
  - It involves saving raw data, simulated data, analysis code, conditions and comprehensive documentation.
  - This approach ensures that methods and context are fully understood, allowing others to reliably replicate the analyses.

Brookhaven National Laboratory

Challenges for Reproduction of Analysis

Some of the challenges that must be overcome for *each* analysis to be preserved:

- Navigating complex analysis methods and workflows,
- Understanding the computing environment, detector and accelerator conditions,
- Preparing Data, analysis artifacts, and documentation for use by non-collaborators.

This requires early and ongoing effort from experts (analyzers, computing, and software) to capture knowledge before it fades.

The efforts in analysis preservation by the PHENIX (and STAR) collaborations have provided valuable insights for the RHIC community.

Different levels of preservation

- Level 3 allows for re-analysis only.
- Most experiments (Zeus, H1, Babar, etc.) have adopted level 4, which allows for simulation and data reconstruction.

Preservation Model		Use Case	
1	Provide additional documentation	Publication related info search	Documentation
2	Preserve the data in a simplified format	Outreach, simple analyses	Outreach, reanalysis
3	Preserve the analysis level software and data format	Full scientific analysis, based on the existing reconstruction	Technical Preservation Projects
4	Preserve the reconstruction and simulation software as well as the basic level data	Retain the full potential of the experimental data	

Data Preservation Levels defined by the Data Preservation in HEP (DPHEP) Collaboration

Aiming at level 4 for RHIC Data and Analysis Preservation

Comprehensive RHIC Data and Analysis Preservation plan

- **Goal**: Ensure the integrity, accessibility, and longevity of RHIC data and analysis post-RHIC operation.
- **Comprehensive strategy**:
  - Engagement with experiments while they are actively analyzing data, integrating DAP into the publication process.
  - Implement standards and best practices as documented, for example, by the Data Preservation in High Energy Physics (DPHEP) collaboration
  - Future users of RHIC data may not be current collaborators, so the DAP strategy should be designed with this in mind.
- **Components of a DAP plan**:
  - Software Preservation: Maintain operability of older software using containerization.
  - Knowledge Management: Capture and preserve knowledge through detailed documentation.
  - FAIR Principles: Ensure data is Findable, Accessible, Interoperable, and Reusable.
  - Dedicated Portal: Provide comprehensive access to RHIC data and analyses
- **Continuous Improvement**:
  - DAP will evolve through collaboration with the RHIC user community.
  - Each experiment is different
- Note: bit and functional preservation (storage and infrastructure) is not covered (yet) by this project

A BNL Program Development for supporting RHIC Data and Analysis Preservation was approved for FY25.

Bit and functional preservation

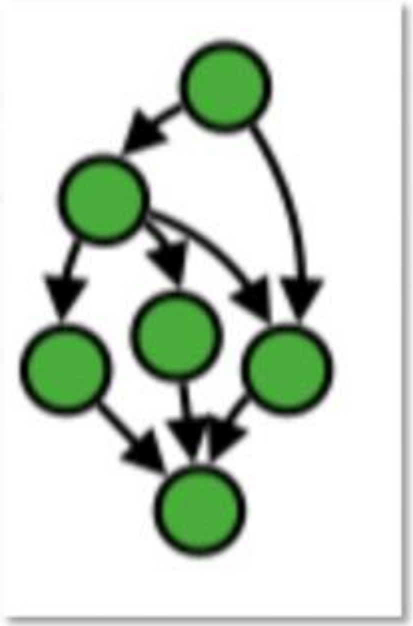
- **Bit Preservation** ensures that digital files remain unaltered over time, preventing data loss and corruption.
- **Functional Preservation** ensures that data remains accessible and usable over time.

Both types of preservation are essential for a successful Data Analysis and Preservation strategy.

This means that data archives (tape systems, media, and software) and some computing infrastructure must be maintained and remain accessible.

Software Preservation

- **Challenge**:
  - Maintaining older software across diverse operating systems and hardware architectures.
- **Solution**:
  - Use containerization technologies to encapsulate analysis software along with its dependencies.
  - Preserve workflow orchestration of analysis steps to ensure continued operability on future platforms.



While container technology ensures the portability of software environments, it remains confined to the original environment and technologies. Software must be modernized and adapted for successful long-term preservation and analysis.

Knowledge Management

- **Importance**:
  - Preserve expertise through detailed accurate documentation before it fades over time.
- **Implementation**:
  - Document data formats, analysis pipelines, and scientific rationales behind specific analyses.
  - Record and document experimental conditions
  - Control and document software versions.
  - Use version control systems like Git to track changes and ensure traceability.

This is crucial and difficult

Implementing FAIR Principles

- **Framework**:
  - Ensure data is Findable, Accessible, Interoperable, and Reusable (FAIR).
- **Implementation**:
  - Use standardized metadata and persistent identifiers (PIDs) to enhance discoverability.
  - Include comprehensive metadata for each dataset, describing its content, context, and provenance.
  - Assign unique IDs and metadata to each analysis for advanced search capabilities.

A dedicated DAP Web Portal

Organize information for each analysis by providing links to Data, Code, Workflow, Plots, Figures, and Documentation

- Single Discovery Entry:
  - Access various data and information about RHIC analyses.
  - Perform simple and complex searches for data.
- User-Facing:
  - Provides access to data analysis tools and pertinent information.
  - Links to repositories with detailed datasets and analyses.
  - Accessible through Federated Identity mechanisms initially for RHIC collaborators only
- Incorporating an AI assistant (Chatbot) leveraging Large Language Models:
  - To manage large volumes of information specific to RHIC.
  - Provide code snippets, facilitate data access, and answer complex questions.
  - Developed through collaborative learning with experts from the collaborations.
  - STAR has already a prototype.

Summary

- Data and Analysis Preservation (DAP) at RHIC is vital for scientific integrity and accessibility.
- A comprehensive DAP at RHIC was proposed, establishing solid foundation for a preservation program which will require long-term support.
- Implementing successful RHIC Data and Analysis Preservation will require significant effort and commitment from the entire RHIC community, including collaborations and the Lab.
- A task force with all RHIC experiments would boost collaboration and efficiency.