Data and Analysis Preservation: the PHENIX Perspective

M.Potekhin for PHENIX Collaboration

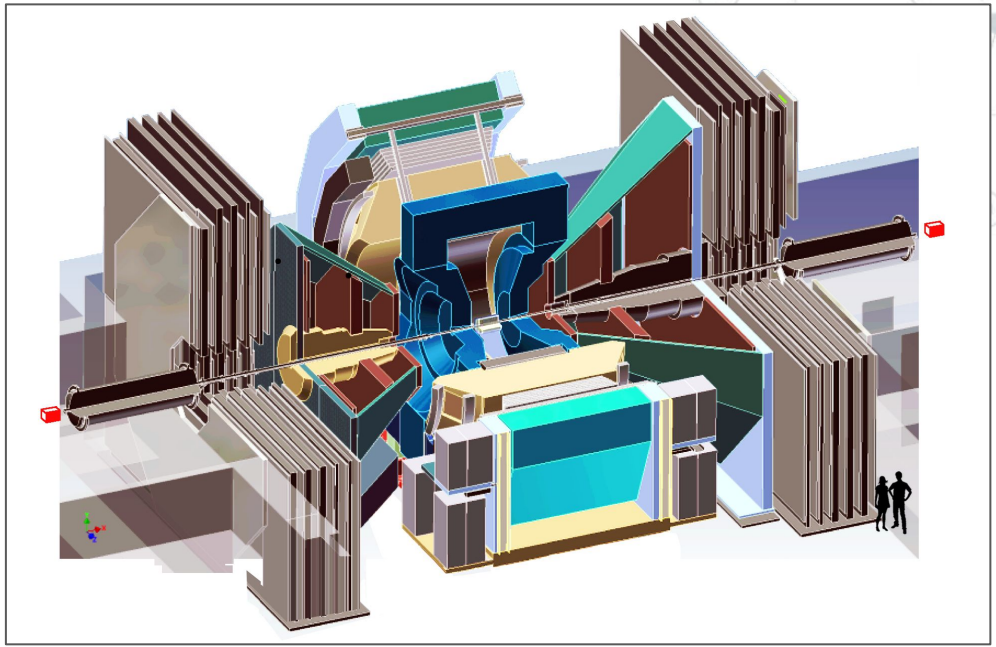
RHIC DAP Roundtable, Jan 16th 2025

Some links of interest

- PHENIX has started its Data and Analysis Preservation (DAP) effort in 2019, and periodic status updates have been presented in the previous DPHEP (CERN) meetings, PV2023 and two ACAT conferences (2021 and 2024). Useful links:
  - https://indico.cern.ch/event/1043155/timetable/#9-bnl-rhic
  - https://iopscience.iop.org/article/10.1088/1742-6596/2438/1/012020
  - https://doi.org/10.5281/zenodo.7905555
- This presentation is a brief summary of our experience and “lessons learned” in this area.

PHENIX

- RHIC at Brookhaven National Laboratory – “Relativistic Heavy Ion Collider” – is one of only two operating heavy-ion colliders (and the only one in the US)
- PHENIX – “Pioneering High Energy Nuclear Interaction eXperiment”: a large, complex detector with a considerable physics reach
- Data taking finished in **2016**, many data analyses are still ongoing and results are being published



DAP in PHENIX: taking stock in 2019

- Since the end of the data taking, the PHENIX member base has been diminishing, leading to less available effort and loss of know-how
- Before 2019, DAP was not prioritized – the right decision to do so was made at the right time – a working group was formed
- Web-based information systems: both the technology platform and the content were quickly becoming obsolete (more detail in the next slides)
- Knowledge management – largely confined to the “analysis notes” approach, which in practice is rarely sufficient to meet the goals. Documents stored in an in-house database with limited capabilities
- HEPData effort – largely stalled back in 2019 – this changed

PHENIX web presence in 2019

- The main website was showing signs of age, in terms of both design and capabilities, and also the content
- Updates (such as PHP and other components) were lagging, leading to eventual flagging and blocking by Cybersecurity – creating a serious hurdle for the Collaboration
- Information related to data taking hosted separately (e.g. trigger masks) and cluttered with materials no longer relevant
- cf. the run catalog did not exist as a proper file or a web page

PHENIX web presence – solution

- The website has been completely redesigned, based on the “static website generator” technology
- Fast and secure
- Very little maintenance required
- The key to its usefulness: Jekyll+Markdown
  - Structured data storage (YAML), providing capabilities close to a DB in some cases, without the actual DB
  - Flexible macro language
  - Ability to pick a template of choice, i.e. content and presentation layers are completely separate
- Close to ideal for the long-term DAP environment

Other Solutions

- Publication data (data points)
  - Substantial progress has been made in 2023, in publishing the data on HEPDATA (vast majority of all papers, ~214)
  - This has replaced the pre-HEPData storage of the PHENIX publication data in text files, in arbitrary formats
- Knowledge management and research document management
  - Leveraged the Zenodo platform at CERN, complemented with a list of curated keywords with direct links from the PHENIX website
  - ~700 PHENIX items published on Zenodo
  - Close to ideal for storage and management of any type or write-ups, PhD theses, conference presentations, all indexed and findable
- Software environment preservation
  - Version control, containerization, REANA

Zenodo + the Website = Synergy

- The website https://www.phenix.bnl.gov/:
  - Built with a static website generator
  - Fast and secure
  - Extensive use of YAML for structured data
- Curated keywords: 315 at the time of writing, automatically translated into functional Zenodo links to the PHENIX materials committed to this platform

The website: the catalog of conference presentations, with auto-generated links to Zenodo

Conferences (132 items)

Keyword	Description
acat24	ACAT 2024
aum16	RHIC & AGS Annual Users Meeting (2016)
aum17	RHIC & AGS Annual Users Meeting (2017)
aum18	RHIC & AGS Annual Users Meeting (2018)
aum19	RHIC & AGS Annual Users Meeting (2019)
aum20	RHIC & AGS Annual Users Meeting (2020)
aum21	RHIC & AGS Annual Users Meeting (2021)
aum22	RHIC & AGS Annual Users Meeting (2022)
aum23	RHIC & AGS Annual Users Meeting (2023)
aum24	RHIC & AGS Annual Users Meeting (2024)
charm21	10th International Workshop on CHARM Physics
cipanp18	Conf. on the Intersections of Particle And Nuclear Physics (2018)
cipanp22	Conf. on the Intersections of Particle And Nuclear Physics (2022)
cpod17	Critical Point and Onset of Deconfinement (2017)
cpod18	Critical point and Onset of Deconfinement (2018)
cpod22	Critical Point and Onset of Deconfinement 2022
dis17	Deep Inelastic Scattering (2017)
dis18	Deep Inelastic Scattering (2018)
dis19	Deep Inelastic Scattering (2019)
dis21	Deep Inelastic Scattering (2021)
dis22	Deep Inelastic Scattering (2022)
dis23	Deep Inelastic Scattering (2023)
dis24	Deep Inelastic Scattering (2024)
dnp19	DNP (2019)

The website: physics keywords, as functional links

Physics (98 items)

Keyword	Description
3he+au	Helium3-on-gold collisions
anisotropy	Anisotropy
asymmetry	Asymmetry
au+au	Gold-on-gold collisions
azimuthal	Azimuthal
b-fraction	fraction of b-quarks
b-meson	B meson
backward-rapidity	The backward kinematic region
binary scaling	Binary scaling
bose-einstein	Bose-Einstein statistics
bottom	Particles containing the b-quark (bottom)
centrality	Centrality characteristic of the collision
cgc	Color Glass Condensate (type of matter)
charm	Particles containing the c-quark (charm)
charmonium	Meson containing a c-quark and its antiparticle
cnm effects	Cold Nuclear Matter effects
correlations	Various types of correlations
cronin effect	Cronin effect
cross section	Cross section (as it applies to scattering)
cu+au	Copper-on-gold collisions
cu+cu	Copper-on-copper collisions
cumulant	Cumulant
d+au	Deutron-on-gold collisions
d-meson	D meson
dca	Distance of Closest Approach
dielectron	A pair of electrons
dilepton	A pair of leptons
dimuon	A pair of muons produced in a collision
direct photon	Direct photons produced in nuclear collisions
drell-yan	Drell-Yan type of process

More complex (multi-keyword) queries can be constructed directly on the Zenodo website if necessary.

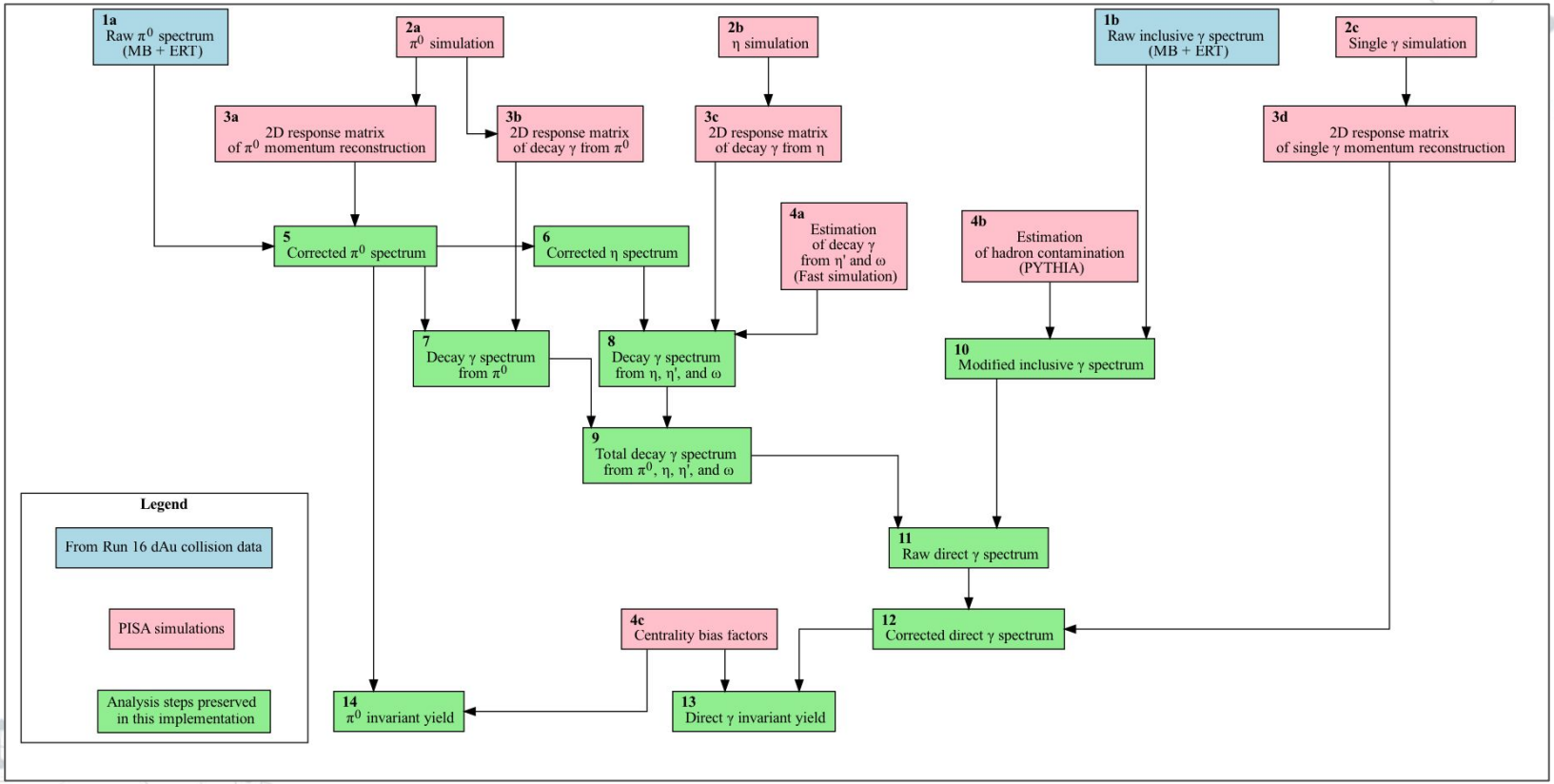
Experience in Analysis Preservation

- To state the obvious: analysis preservation is only possible with full commitment of people involved in that particular analysis, and provided a complete and accurate flowchart of the analysis has been created (the exact format is unimportant)
- Proper version control of the final analysis software is often lacking
- We recently performed a complete preservation of an important analysis (nuclear modification factor in d+Au collisions). One more analysis (dimuon decay of J/psi) is in the pipeline.

An example: direct γ and π^0 analysis

- The main motivation for this study was an initially puzzling result regarding the nuclear modification factor in peripheral d+Au collisions. The background is explained in a PHENIX note: https://doi.org/10.5281/zenodo.8169171
- Ultimately it was determined that the key to understanding this result is to use the correct technique of estimating the centrality of collisions, making use of the Electromagnetic Calorimeter data. This has substantial scientific importance, and hence was chosen for preservation.
- This analysis featured a well documented flowchart and other crucial documentation which made this effort possible.

Direct γ and π^0 analysis: the flowchart



Direct γ and π^0 analysis: documentation (web)

https://www.phenix.bnl.gov/analysis/dAuPi0Photon.html

Direct γ in d+Au collisions

The measurement of γ and π^0 yields in d+Au interactions is important for studying the formation of quark-gluon plasma (QGP) in heavy ion collisions.

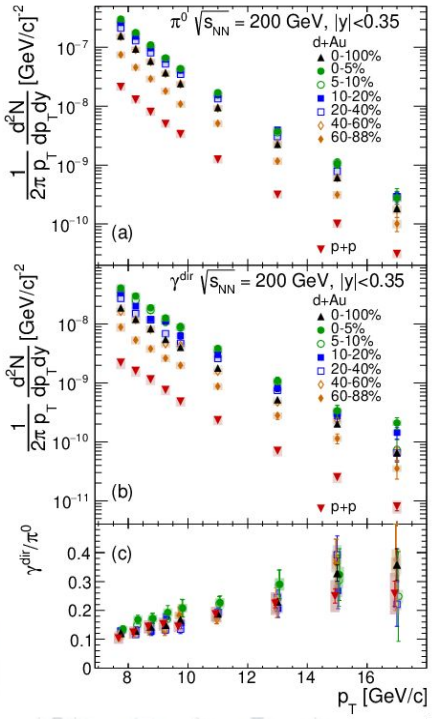
One way to measure QGP formation is by observing jet suppression using the nuclear modification factor R_{AB}, which compares the yield of a particle (in this case, the π^0) observed in AB is the same as that observed in p+p. If R_{AB} is less than one, then the yield in AB is suppressed, and if it is greater than one, then it is enhanced.

For a more detailed explanation that includes the motivation and physics background, please refer to this write-up: DOI 10.5281/zenodo.8169171.

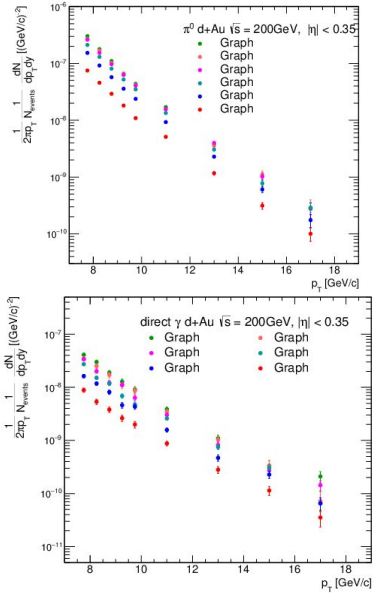
A detailed, step-by-step description of the analysis procedure, with references to the tagged blocks in the flowchart.

Validation of the procedure

Published



REANA run



A test run of the preserved analysis by a person (a computer science student who is not a PHENIX member), based solely on documentation, produced the correct published result.

First publication from RHIC on the topic of the preserved analysis:

https://arxiv.org/abs/2408.12072

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

REANA + CVMFS

- REANA is a helpful platform for DAP, conducive to containerization and well preserved workflows
- The PHENIX REANA framework dependency on CVMFS exists because of our desire to keep the size of the software image manageable (e.g. under 10GB)
- However, it is reasonable to assume that it is very helpful for a REANA instance to have the CVMFS capability, to be future-proof – likely that other experiments will have a dependency as well
- Initial discussions with the REANA developers indicate a possibility of a solution for platforms like OKD (which by default don't work with CVMFS)

The J/ψ dimuon decay study (a DAP candidate)

- This study is our next candidate for analysis preservation
- PHENIX Publication: Phys. Rev. C 102, 014902
- Scientific significance
  - Heavy quarkonia suppression is one of the hallmark signs of quark-gluon plasma formation
  - Detecting signs of final state suppression due to quark-gluon plasma formation in small systems was done by comparing the nuclear modification measurements at forward and backward rapidity in three different systems: p+Al, p+Au, and ^{3}He+Au
  - The analysis shows that nuclear absorption and gluon anti-shadowing describe the PHENIX data very well at backward rapidity and the suppression is likely not due to quark gluon plasma formation

Wish list

- The RHIC experiments and their Analysis Preservation effort would benefit greatly from an equivalent of the Open Data portal (as it exists at CERN) deployed and available to the RHIC community. It is an efficient way to integrate heterogeneous materials (data, documentation, software) in an accessible package, at any level of complexity
- Open Data is based on Invenio, and Invenio has been supported at BNL for the past few years. Can it be ported here?
- A key component of DAP in general and of platforms such as Open Data is the mass storage, which in case of Open Data is mostly **EOS.** There is currently no **EOS** equivalent in the US in terms of scale and accessibility.

An opportunity

- There are no standards and/or standard platforms for NP experiments in the US to leverage, for Analysis Preservation – as confirmed in a recent “NP software infrastructure” workshop at JLab
- For that reason, there are few possibilities to adhere to the FAIR principles when preserving data and analyses
- Deployment of REANA, an equivalent of the Open Data and EOS storage etc would give BNL the leadership position in this area.

The key issue

- The PHENIX experience in DAP shows that regardless of platforms and technologies deployed and available, having dedicated personnel focusing on this specific work area is absolutely crucial for the success of such efforts
  - Even outside of PHENIX, this is the overwhelming opinion of all major experiments (e.g. at the LHC and CERN in general)
- This is in addition to participating researchers who should embrace these practices.

Summary

- PHENIX has made substantial progress in its Data and Analysis Preservation effort overy the past five years
- The PHENIX web presence has been completely reworked
- In the past two years, most of the publication-related data has been committed to the HEPData portal and this work is ongoing
- Zenodo remains one of the principal PHENIX DAP components with 700 items committed, including indexed presentations from 132 conferences and 170 PhD theses
- REANA-based analysis preservation effort resulted in one analysis preserved in substantial detail and validated against the actual publication plots, and another important analysis is in the pipeline