

Data & Analysis Preservation: Experience in PHENIX

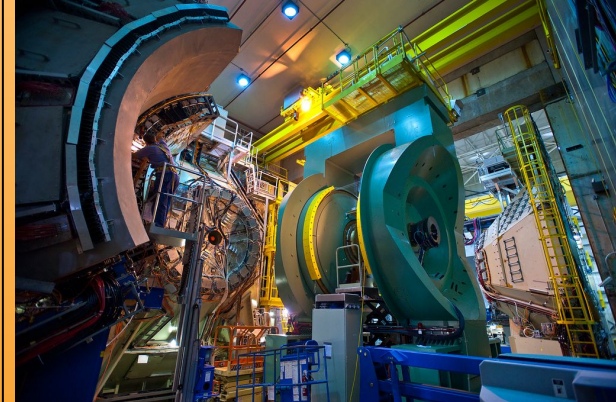
Maxim Potekhin

Nuclear and Particle Physics Software Group (BNL)



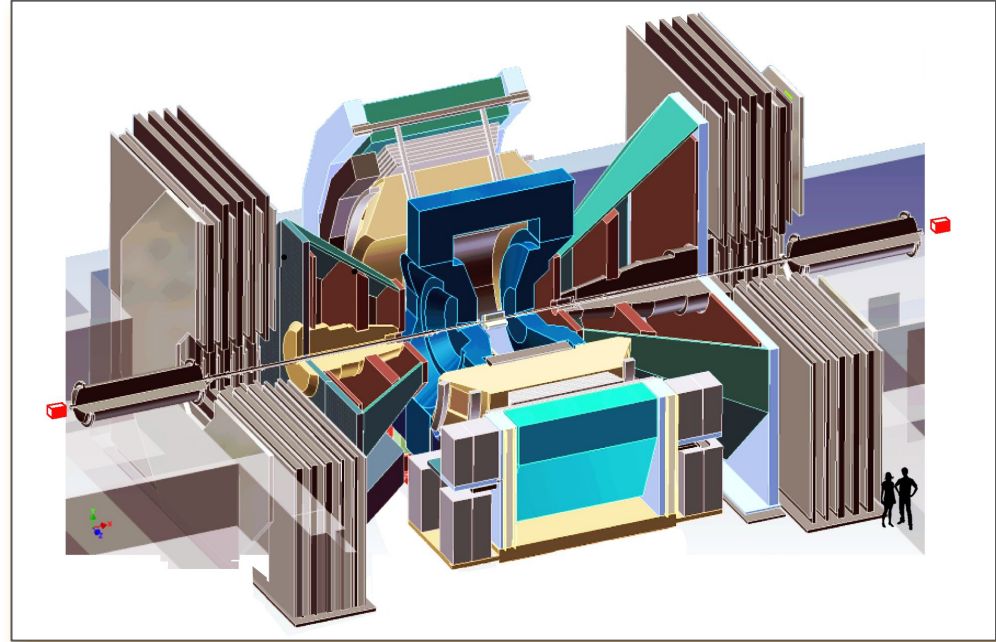
DPHEP Collaboration Workshop 2021

06/23/2021



PHENIX

- “Pioneering High Energy Nuclear Interaction eXperiment”
- One of the two large RHIC experiments
- A large, complex general purpose detector with a considerable physics reach and *complex analyses*



- For more details please see the “PHENIX Collaboration Community” on Zenodo:
<https://zenodo.org/communities/phenixcollaboration/>

PHENIX today

- Data taking finished in 2016 with ~ 24 PB of raw data accumulated
- Active analysis work underway (average ~ 10 articles a year in 2019-2020)
 - Total of >240 published papers + conference contributions (total ~ 1200 items), ~ 165 PhD theses and counting

- All current analyses are done using preserved data

- In this presentation we share our experience in the area of Data and Analysis Preservation (*DAP*)

- Effort started in 2019

INSPIRE

PHENIX

Literature Authors Jobs Seminars Conferences More...

1,365 results

Date of paper

Number of authors

Exclude RPP

Document Type

Author

Probing gluon spin-momentum correlations in transversely polarized protons through midrapidity isolated direct photons in $p + p$ collisions at $\sqrt{s} = 200$ GeV

PHENIX Collaboration • J.A. Acharya (Georgia State U.) et al. (Feb 26, 2022)

e-Print: 2102.11985 [hep-ex]

Highlights from the PHENIX Collaboration

PHENIX Collaboration • Megan Connors (Georgia State U. and RIKEN BNL) for the collaboration. (Jan, 2021)

Published in: Nucl Phys A 1005 (2021) 121925 • Contribution to: Quark Matter 2019

Signatures of collectivity and flow in small systems observed by PHENIX

PHENIX Collaboration • Seoyoung Hong (Korea U.) for the collaboration. (Jan, 2021)

Published in: Nucl Phys A 1005 (2021) 122008 • Contribution to: Quark Matter 2019

Modification of hadron productions in small systems observed by PHENIX

PHENIX Collaboration • N. Mikkelsen (DESY) for the collaboration. (Jan, 2021)

Published in: Nucl Phys A 1005 (2021) 121878 • Contribution to: Quark Matter 2019

Quark flavor dependence of particle flow in nucleus-nucleus collisions measured by PHENIX

PHENIX Collaboration • Takahito Todoroki (RIKEN BNL) for the collaboration. (Jan, 2021)

Published in: Nucl Phys A 1005 (2021) 121960 • Contribution to: Quark Matter 2019

PHENIX J/ψ Measurements in $p + A$, $p + Au$, and $^3\text{He} + Au$ collisions

Challenges (applicable to PHENIX)

If there is one lesson in this story it is the need to take a “holistic approach” – data without the software is often useless, as is software without build and verification systems and/or necessary additional data (alignment, calibration, magnetic field maps etc.) These are typically stored separately and involve distinct services that evolve on independent timescales and with lifetimes typically much shorter than the period for which the corresponding “data” needs to be preserved.

<https://doi.org/10.5281/zenodo.2653526> “Software Preservation and Legacy issues at LEP” (J.Shiers)

No matter what preservation tools are developed that might enable reuse of software, analysis techniques, and data, if they are not conceived from the beginning as an integral part of the standard frameworks, retrofitting will be nearly impossible.

<https://arxiv.org/abs/1810.01191> “HSF White Paper: Data and Software Preservation to Enable Reuse”

The role of the facility



- DAP universally depends on *continuity of services and expertise* provided by the facility
 - cf. the previous slide
- This is especially true for PHENIX: BNL SDCC is its only functioning computing site.
- In addition to bit preservation (mass storage) the facility provides software builds and provisioning capabilities (including containers, CVMFS etc), databases and more.
- Any planning of DAP must include facility involvement over the relevant time period.

DPHEP: BNL Participation

- BNL is a member of the DPHEP Collaboration
- SDCC (BNL) is an active DPHEP partner on the facility side, participating in DAP technology development and testing.
- BNL “Nuclear and Particle Physics Software Group” and PHENIX members participated in the DPHEP Workshop at CERN in 2019 and continue to be actively engaged with and receive guidance from DPHEP

PHENIX: Challenges of Knowledge Management

- Need to keep records of software provenance, dependencies, configuration, use etc
- Software preservation \neq Analysis preservation
- Keep track of “data artifacts” such as conditions-type data which may be produced for the purposes of a particular analysis and depend on details known mostly to the people involved in this analysis (misc. cuts, maps, lists, numerical constants in macros etc)
- There is a legacy solution which is a requirement to record such info in a dedicated section of the “Analysis Note” which must accompany every paper, but in reality its efficacy is variable
- Hard to provide continuity of know-how as people move on

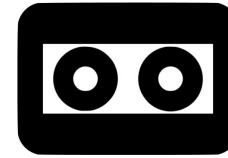
PHENIX: Legacy web infrastructure

- Information was spread across a few legacy web resources - the software, detector and subsystem information and other documentation
- Information was diluted with items once relevant for PHENIX but no longer current or aligned with its current and future needs
- PHP-based proprietary information systems e.g. document database (papers, talks, theses), numerical data archive etc became difficult to upgrade, maintain and keep secure - and experiences outages

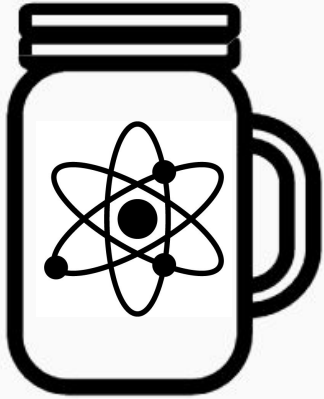
PHENIX: Software challenges

- Over the years, portability of the core software build procedures was largely lost
 - Build and configuration are specific and coupled to the computing site (BNL)
 - Situation not unique to PHENIX
- Due to compatibility issues most of the software is still built in the i686 environment
 - This does mean extra software packages that need to be installed on modern Scientific Linux
 - Getting hold of certain gcc/OS combinations etc may be a challenge sometimes
- ROOT5 still widely used (and is default) due to many legacy macros
 - Dependencies must also be addressed for the 32-bit build
- PHENIX leveraged Singularity to run production in a containerized environment
 - Motivated by reproducibility, keeping SL6
 - ...with the caveat that the software stack is in AFS - not suitable for REANA

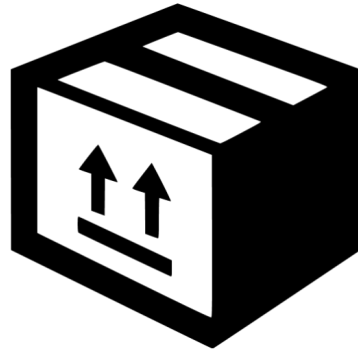
The DAP Strategy in PHENIX



Bit preservation



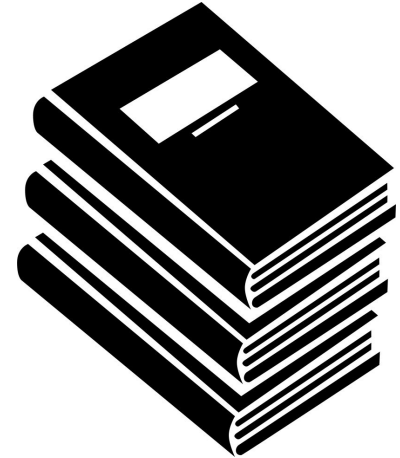
Analysis capture



Containers



Updated
Web-based
documentation



Modern repositories for
research materials

The new website: <https://www.phenix.bnl.gov/>

PHENIX

PHENIX, the Pioneering High Energy Nuclear Interaction eXperiment, is the largest of the four experiments that have taken data at the Relativistic Heavy Ion Collider. Data-taking was finished in 2016 and the PHENIX Collaboration is now analyzing large data samples previously collected, prioritizing those with a unique physics reach.

This website has been created by the Collaboration to support its Data and Analysis Preservation effort. Materials are collected from legacy web resources, curated and systematized for placement on this site. New content is created with the specific purpose of aiding in the long-term analysis preservation. This effort leverages best practices and tools developed in the High Energy and Nuclear Physics communities such as Zenodo, HEPData and Open Data portals.

Temperature

LHC

PHIC

Quark-Gluon Plasma

Hadrons

Atomic nuclei

Neutron stars

Baryon Density

PHENIX members: please examine [the list of work items](#) and [let us know](#) if you can help.

Site built at 2021-06-16 18:58:14 -0400

Brookhaven National Laboratory

PHENIX

U.S. DEPARTMENT OF ENERGY

The website functionality and design

- Links to various (and new) PHENIX resources are provided and managed on the site:
 - Zenodo, HEPData, OpenData
 - InspireHEP
 - GitHub, Docker Hub
 - REANA
 - Technical notes and descriptions of the detector subsystems, run history etc (hosted locally)
- The website is effectively replacing legacy web resources
 - With a lot of functionality moved to cloud platforms listed above
- Design goals - ease of long term maintenance, performance and security
- We are using a static site generator (Jekyll) to achieve these goals
 - Inspired by the HSF website design
 - Development version is hosted on GitHub pages
 - Production version is hosted at BNL
 - Helper macros developed (in Liquid) to make content creation and management easier

HEPData and OpenData

- Level 1: Data Products used in publications.
 - Such as data points and errors used in plots, in numeric format
 - cf. the “HEPData” portal: <https://www.hepdata.net/>
- Level 2: Special Purpose Datasets for Education and Outreach.
 - Select datasets + virtualized or otherwise portable analysis software + documentation
 - cf. the “OpenData” portal: <https://opendata.cern.ch/>
- Level 3: Reconstructed Open Data; may be released in future
 - Implies a more complex analysis environment than in Level 2
 - Requires adequate software and computing infrastructure to be properly used
- Level 4: Raw Data. Preserved, but not considered useful for release.



PHENIX on HEPData

The screenshot displays the HEPData search interface. At the top, the search bar contains 'phenix' and the results are filtered to 'PHENIX'. The page shows 47 results, with the first three displayed. Each result includes the title, authors, publication reference, and a brief abstract. The first result is 'Centrality dependence of charged particle multiplicity in Au - Au collisions at $\sqrt{s(NN)^{1/2}} = 130$ GeV' by The PHENIX collaboration. The second is 'Dilepton mass spectra in p+p collisions at $\sqrt{s^{1/2}} = 200$ GeV and the contribution from open charm' by The PHENIX collaboration. The third is 'Measurement of the mid-rapidity transverse energy distribution from $\sqrt{s(NN)^{1/2}} = 130$ -GeV Au + Au collisions at RHIC' by The PHENIX collaboration. On the left side, there are filters for Date, Collaboration (PHENIX), Subject_areas, Phrases, Reactions, and Observables. A bar chart shows the distribution of results over time from 2001 to 2020.

HEPData submissions mandated for all new publications

Revisiting older publication materials as time permits

Using GitHub for material development, support of team effort

PHENIX OpenData entry - the first for a US-based experiment

The screenshot shows the CERN OpenData portal interface. At the top, there is a search bar with the text 'open data CERN' and a search icon. Below the search bar, there are navigation links for 'Help' and 'About'. The main content area is divided into several sections:

- Filter by type:** Includes 'include on-demand datasets' (unchecked), 'Dataset' (checked), and 'Derived' (unchecked).
- Filter by experiment:** Lists experiments with their respective counts: ALICE (26), ATLAS (127), CMS (2694), LHCb (12), OPERA (910), and PHENIX (1, selected).
- Filter by file type:** Lists file types: 'c' (1), 'pdf' (1), and 'root' (1).
- Filter by event number:** Lists event number ranges: '0-999' (0), '1000-9999' (0), '10000-99999' (0), '100000-999999' (0), '1000000-9999999' (1), and '10000000--' (0).

The search results section shows 'Sort by: Best match' and 'asc.', and 'Display: detailed' and '20 results'. It states 'Found 1 result.' and displays a record titled 'Examples of basic analysis techniques for neutral meson and photon data from the PHENIX detector'. The record description reads: 'This record contains datasets from the Electromagnetic Calorimeter (EMCal) of the PHENIX detector. It aims to present a few basic techniques of identifying π^0 -s and photons using that device. The data...'. Below the record, there are buttons for 'Dataset', 'Derived', and 'PHENIX'. At the bottom of the page, there are logos for ALICE, ATLAS, CMS, LHCb, OPERA, and PHENIX, along with the CERN logo and copyright information: '© CERN, 2014–2021 · Terms of Use · Privacy Policy · Help'.

- Contents of the package:
 - Derived data (Ntuples)
 - ROOT macros
 - Detailed instructions (PDF)
- Subject area:
 - Analyses based on the EM calorimeter data
- Thanks to the CERN team for helping this happen!
- Plan: to add more instructive items of this type

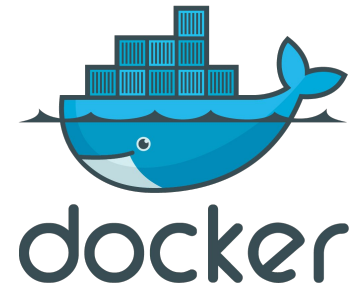
Zenodo@CERN - the PHENIX community

<https://zenodo.org/communities/phenixcollaboration>

- ~400 PHENIX items, uploads ongoing
- Branded, curated, discoverable, DOI'd
- Well-suited for long-term preservation
 - Also works well for current activity: theses, analysis tutorials, conferences etc
 - 99% percent of document storage is outsourced here from the PHENIX website, taking full advantage of Zenodo search capabilities
- Indexed
 - Keywords are managed and linked on the PHENIX website
- +elastic search capability

The screenshot displays the Zenodo interface for the PHENIX Collaboration community. At the top, the Zenodo logo is on the left, a search bar in the center, and 'Upload' and 'Communities' buttons on the right. A user profile 'phenix-dap-l@lists.bnl.gov' is visible in the top right corner. The main heading is 'PHENIX Collaboration'. Below it, a 'Recent uploads' section features a search bar and a list of three items. Each item includes a date, version, and status (Thesis, Presentation, Open Access), a title, author name, a brief description, and an upload date. The first item is a thesis by Wong, Cheuk-Ping about π^0 -hadron correlations. The second is a presentation by Esha, Roli about PHENIX measurement of system size dependence. The third is a presentation by Wong, Cheuk-Ping about jet modifications. On the right side, there is a 'New upload' button, the PHENIX logo, and a description of the community's purpose. Below that, it lists the curator (PhenixCollaboration), the creation date (May 18, 2020), and the harvesting API (OAI-PMH Interface). At the bottom right, there is a section asking if the user wants their upload to appear in the community, with a button to click.

Capturing the Software Environment



- For most of the PHENIX software the build is not portable
- Containerization offers a partial solution to this problem
 - Also opens the possibility to use REANA (next slide)
- Work is currently underway to create images of the analysis software environment using two different methods
 - Deployment of PHENIX libraries and dependencies on CVMFS for REANA applications
 - Creating a custom image by packaging most relevant libraries
- Also created images to preserve legacy ROOT5 versions to ensure compatibility with older macros
- We are using GitHub to manage Dockerfiles, Docker Hub for image delivery and also a private Docker registry at BNL to provision software to REANA








REANA

- **Deployed at BNL**, with PHENIX team currently on the learning curve, running analysis macros
- Synergy with the recent EIC effort
- Both storage and CPU can be scaled up if resources are made available
- There is interest in running complete final stages of analyses in this environment
- Tutorials prepared for the PHENIX School'21 (session took place today!)

reana Home Examples Get Started Documentation News Roadmap Contact Blog

reana

Reproducible research data analysis platform

Flexible	Scalable	Reusable	Free
Run many computational workflow engines.	Support for remote compute clouds.	Containerise once, reuse elsewhere. Cloud-native.	Free Software. MIT licence. Made with ❤ at CERN.
 COMMON WORKFLOW LANGUAGE	   kubernetes slurm	 	

DAP: Data and Analysis Preservation in 2020s

- In the past few years, DAP has gained an increased prominence in the scope of effort of major High Energy and Nuclear Physics (HEP/NP) experiments, driven by the policies of the funding agencies as well as realization of the benefits brought by DAP to the science output of many projects in the field.
- Platforms like REANA, Zenodo/Invenio, OpenData, HepData form a fairly complete DAP ecosystem, placing solid DAP capability within the reach of most experiments and reducing the need for in-house development
- Observation: pursuit of reproducibility of calculations is not limited to HEP and other sciences but is an integral part of many industrial projects, leading to the establishment of techniques and practices we can learn from
- Knowledge Management is perhaps the central part of DAP, beyond the software and infrastructure preservation (as challenging as these items are)

DAP evolution: from tape archives to DevOps

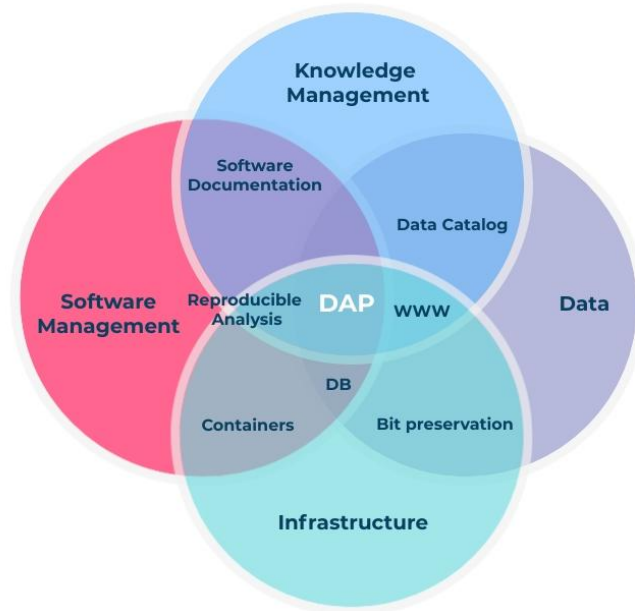
- In the past, there was an assumption that while DAP required investment and effort upfront, any benefits coming from it could only be realized in the long-term, thus complicating adoption of DAP practices. This may no longer be true.
- The technology landscape has changed.
- Many aspects of DAP overlap with modern practices of software development, management and packaging which have immediate impact
 - Version control and code organization
 - Containerization, CI, testing and validation
 - More generally, “software sustainability”
- Knowledge Management (KM) is a core component of DAP but in fact not exclusive to it
 - KM is conducive to efficient knowledge transfer which brings about efficiencies. Consider onboarding new members of a collaboration, bringing graduate students up to speed etc
- Reproducibility is a key factor in creating high-quality scientific output and therefore has both near-term and long-term benefits.

DAP practices have the potential to enhance quality of the science output in near term by helping ensure reproducibility

DAP focus on knowledge management is conducive to efficient knowledge transfer within the collaboration and across projects

Software management, packaging and containerization facilitates deployment

Modern digital repositories create efficient document management solutions on every time scale (cf. the use of Zenodo in both PHENIX and EIC)



Lessons learned

- *DAP: plan and start early*
 - The effort will pay for itself by increasing overall productivity of the experiment
 - PHENIX is fighting an uphill battle here due to a late start
- Avoid building in-house information systems, there are many tools available
 - State-of-the-art services such as Zenodo, OpenData, HEPData, REANA, Inspire etc cover a vast majority of the experiments' needs
- Containerization solves many of the challenges of capturing the software environment
 - Use it!
- Create websites for the long haul (static site generation works well)
 - Avoid platforms that will require updates and maintenance in the long term e.g. Drupal
 - Do not fragment web space across multiple institutions
- Prioritize analyses for preservation as effort is always limited

Final thoughts

- DPHEP Collaboration plays a key role in how PHENIX addresses DAP
 - Thank you!
- What is the expected lifetime of platforms like Zenodo, HEPData etc?
- We are open for DAP collaboration with other projects
 - Our experience will help experiments at BNL (sPHENIX, EIC etc)
- OpenData has potential beyond the LHC
- New experiments need to be made aware that DAP brings benefits not only in the long term, but also fairly immediately - computing models should account for that
- Wider adoption of REANA is a worthy goal