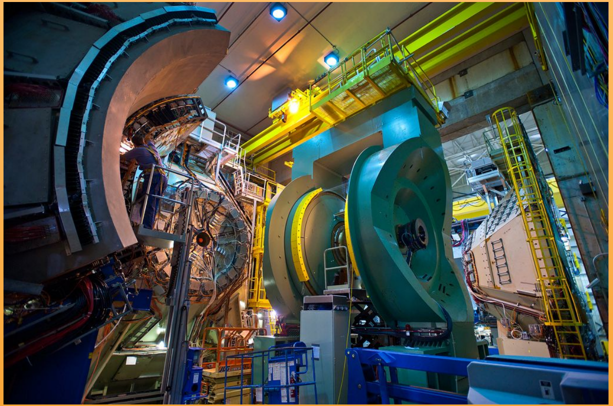
Data & Analysis Preservation

Maxim Potekhin
Nuclear and Particle Physics Software Group (BNL)

Brookhaven National Laboratory

Phenix School 2021



PHENIX

What is DAP?

- The goal of the **Data and Analysis Preservation** (DAP) is to maintain the capability of experiments to reliably perform analyses over a long period of time, thus protecting and leveraging the significant investment of the funding agencies and the science community.
- Retaining data (i.e. so-called *“bit preservation”*) only makes sense if the analysis expertise and necessary software and infrastructure elements are equally well preserved.
- Being able to *access and process* data previously collected opens up opportunities to apply novel analyses techniques, test new models and make corrections if necessary.
- DAP also has important outreach and educational aspects.
  - cf. the “Open Data” policies of major LHC experiments
- PHENIX is engaged in the international DAP community and aims to utilize best in class, well supported platforms for this task.

Challenges according to experts (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"

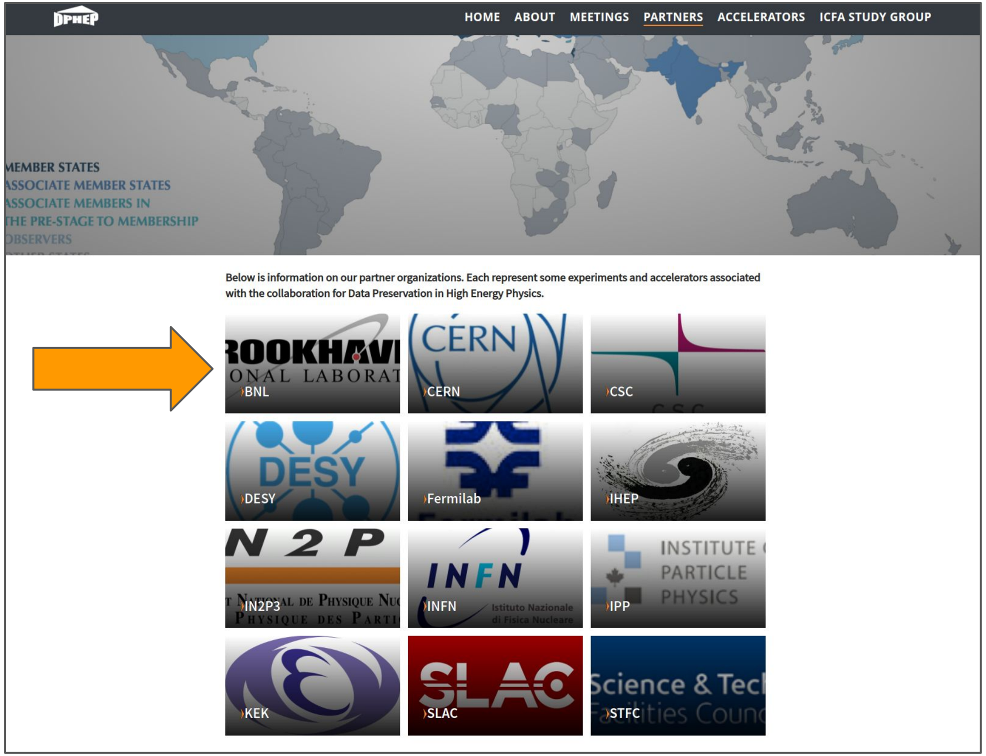
The role of the facility

BROOKHAVEN NATIONAL LABORATORY | Scientific Data and Computing Center

- 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.

The DPHEP Collaboration https://dphep.web.cern.ch/

International Collaboration for Data Preservation and Long Term Analysis in High Energy Physics



Below is information on our partner organizations. Each represent some experiments and accelerators associated with the collaboration for Data Preservation in High Energy Physics.

DPHEP: BNL Participation

- BNL is a member of the DPHEP Collaboration which was formed at CERN ca. 2013 *“The collaboration aims to create a natural forum for the high energy physics community to foster discussion, archive consensus, and transfer knowledge on technological solutions and the diverse governance applying to the preservation of data, software, and know-how in the high energy physics community.”*

- SDCC (BNL) is an active DPHEP partner on the facility side, participating in DAP technology development and testing. This gives BNL and the RHIC/EIC community optimal access to the state-of-the-art methodologies and tools.

- PHENIX members participated in the DPHEP Workshops at CERN in 2019 and 2021 - the latter takes place this week - and continue to be actively engaged with and receive guidance from DPHEP

PHENIX: DAP Challenges, categorized

PHENIX: Challenges of Knowledge Management

- Need to keep records of software provenance, dependencies, configuration, use etc
  - cf. the exact sequence of analysis macros
  - Parameters, arguments, provenance and location of “dead maps” etc
- **Software preservation ≠ 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 requirement to record such info in a dedicated section of the “Analysis Note” which must accompany every paper
  - e.g. verbal description of the procedure and location of macros in HPSS
  - ...but in reality its efficacy is variable, some notes are better than others
- 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 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 we experienced outages

PHENIX: Software challenges

- Over the years, portability of the core software build procedures was largely lost
  - This is not unique to PHENIX
  - Build and configuration are specific and coupled to the computing site (BNL)
- Due to certain compatibility issues most of the software is still built in the i686 (32-bit) environment
  - This means extra software packages that need to be installed on modern Scientific Linux
- ROOT5 still widely used (and is default) due to many legacy macros
  - Dependencies must be addressed for the 32-bit build
- PHENIX used Singularity to run production in a containerized environment
  - ...with the caveat that the software stack is in AFS
  - Now running natively in the SL7 environment

The DAP Strategy in PHENIX

Bit preservation

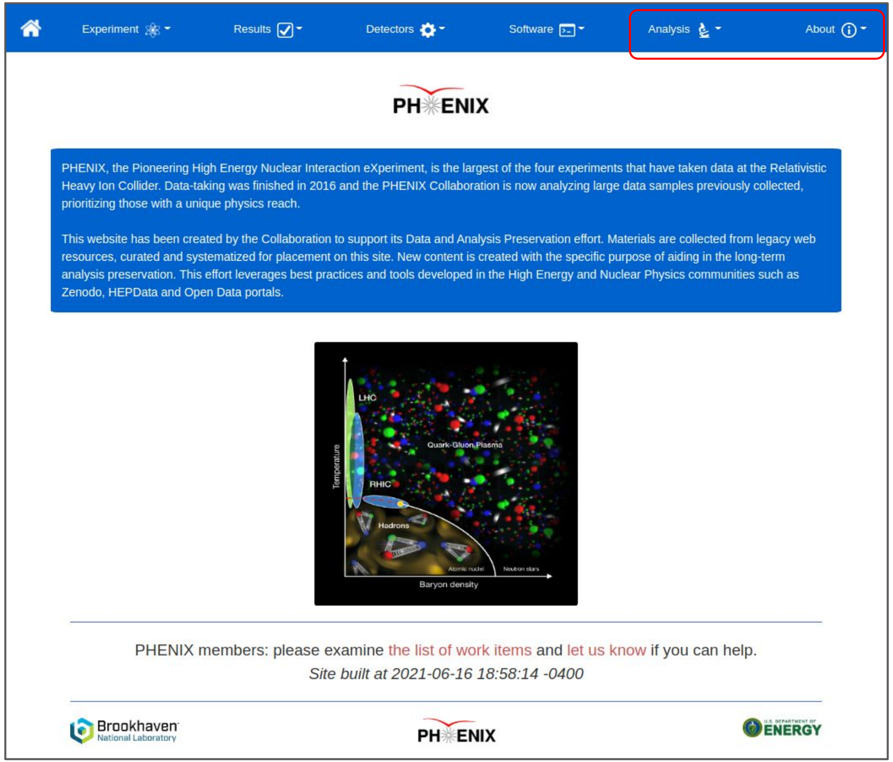
Analysis capture

Containers

Web-based documentation

Modern repositories for research materials

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

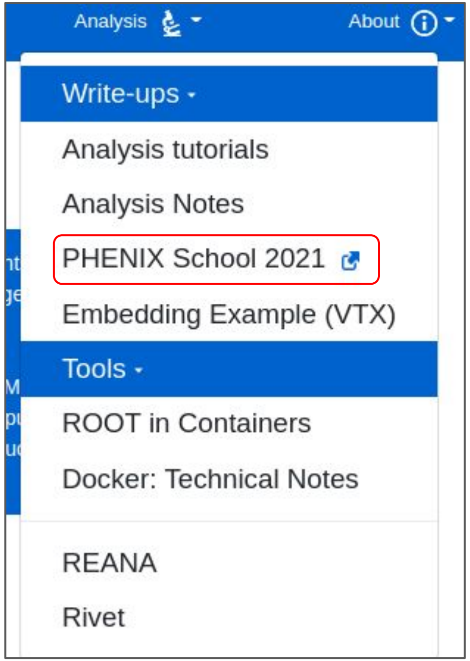


PHENIX, the Pioneering High Energy Nuclear Interaction eXperiment, is the largest of the four experiments that have taken data at the Relativistic Heavy Ion Collider. Data-taking was finished in 2016 and the PHENIX Collaboration is now analyzing large data samples previously collected, prioritizing those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

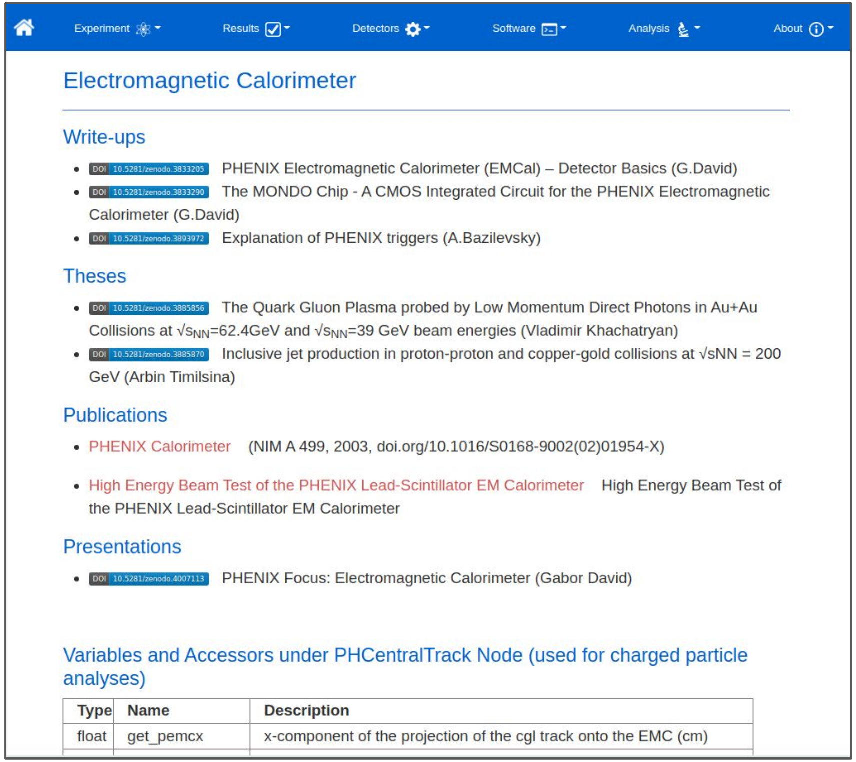
This website has been created by the Collaboration to support its Data and Analysis Preservation effort. Materials are collected from legacy web resources, curated and systematized for placement on this site. New content is created with the specific purpose of aiding in the long-term analysis preservation.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

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



The new website: a typical subsystem page



Electromagnetic Calorimeter

Write-ups

- DOI 10.5281/zenodo.3833205 PHENIX Electromagnetic Calorimeter (EMCal) – Detector Basics (G.David)
- DOI 10.5281/zenodo.3833290 The MONDO Chip - A CMOS Integrated Circuit for the PHENIX Electromagnetic Calorimeter (G.David)
- DOI 10.5281/zenodo.3893972 Explanation of PHENIX triggers (A.Bazilevsky)

Theses

- DOI 10.5281/zenodo.3885856 The Quark Gluon Plasma probed by Low Momentum Direct Photons in Au+Au Collisions at $\sqrt{s_{NN}}$=62.4GeV and $\sqrt{s_{NN}}$=39 GeV beam energies (Vladimir Khachatryan)
- DOI 10.5281/zenodo.3885870 Inclusive jet production in proton-proton and copper-gold collisions at √sNN = 200 GeV (Arbin Timilsina)

Publications

- PHENIX Calorimeter (NIM A 499, 2003, doi.org/10.1016/S0168-9002(02)01954-X)
- High Energy Beam Test of the PHENIX Lead-Scintillator EM Calorimeter High Energy Beam Test of the PHENIX Lead-Scintillator EM Calorimeter

Presentations

- DOI 10.5281/zenodo.4007113 PHENIX Focus: Electromagnetic Calorimeter (Gabor David)

Variables and Accessors under PHCentralTrack Node (used for charged particle analyses)

Type	Name	Description
float	get_pemcx	x-component of the projection of the cgl track onto the EMC (cm)

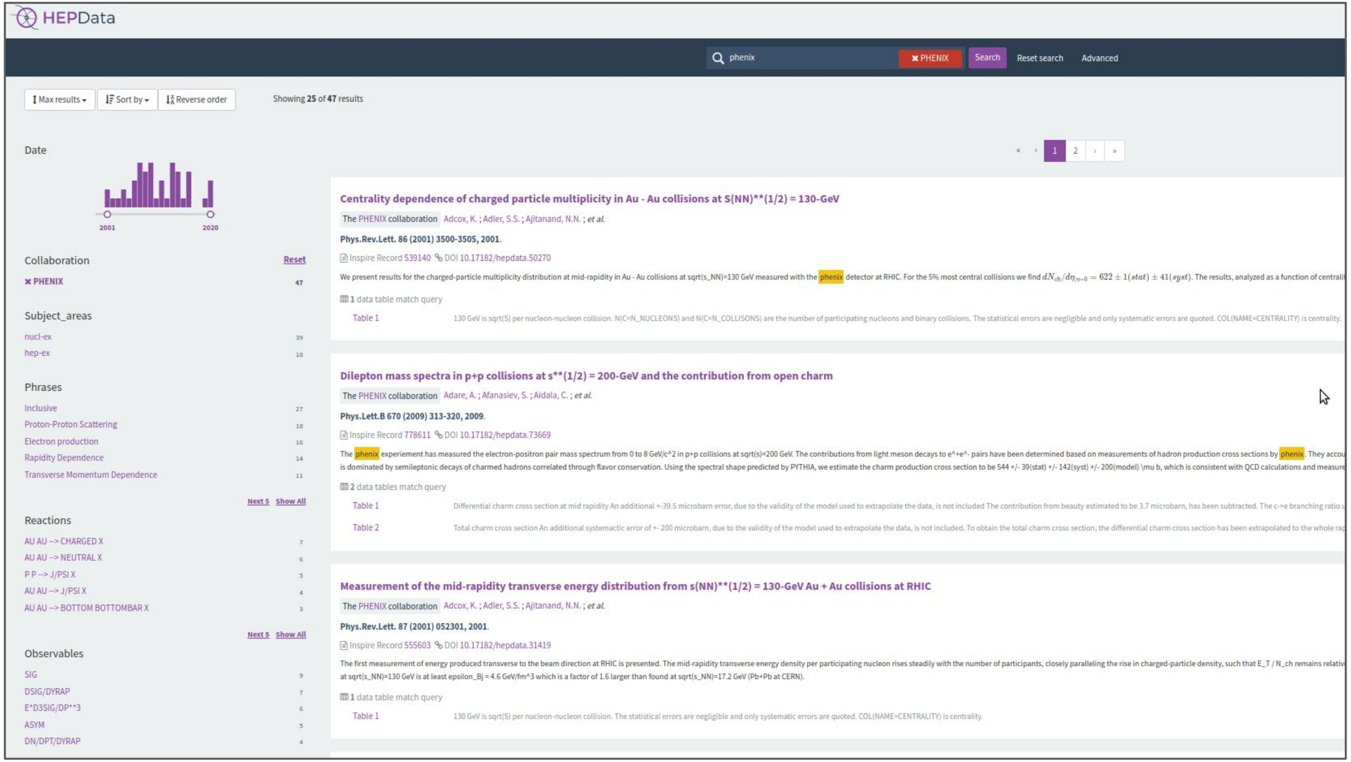
The website functionality and design

- 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 a few legacy web resources
  - With a lot of functionality moved to cloud platforms listed above
  - Some items are hard to migrate (cf. Analysis Note archive), some impossible (CDS)
- Design goals - ease of long term maintenance, performance and security
- We are using a static site generator (Jekyll) to achieve these goals
  - 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
- **Contributions are most welcome!**

Open Data Tiers: HEPData and OpenData portals

- 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



Showing **25** of **47** results

Collaboration	Reset
✖ PHENIX	47

Subject_areas	
nucl-ex	39
hep-ex	18

Phrases	
Inclusive	27
Proton-Proton Scattering	18
Electron production	16
Rapidity Dependence	14
Transverse Momentum Dependence	11

Next 5 Show All

Reactions	
AU AU --> CHARGED X	7
AU AU --> NEUTRAL X	6
P P --> J/PSI X	5
AU AU --> J/PSI X	4
AU AU --> BOTTOM BOTTOMBAR X	3

Next 5 Show All

Observables	
SIG	9
DSIG/DYRAP	7
E*D3SIG/DP**3	6
ASYM	5
DN/DPT/DYRAP	4

Centrality dependence of charged particle multiplicity in Au - Au collisions at S(NN)(1/2) = 130-GeV**

The PHENIX collaboration Adcox, K. ; Adler, S.S. ; Ajitanand, N.N. ; *et al.*

Phys.Rev.Lett. 86 (2001) 3500-3505, 2001.

Inspire Record 539140 DOI 10.17182/hepdata.50270

We present results for the charged-particle multiplicity distribution at mid-rapidity in Au - Au collisions at sqrt(s_NN)=130 GeV measured with the phenix detector at RHIC. For the 5% most central collisions we find $dN_{ch}/d\eta_{|\eta=0} = 622 \pm 1(stat) \pm 41(syst)$. The results, analyzed as a function of centrali

1 data table match query

Table 1 130 GeV is sqrt(S) per nucleon-nucleon collision. N(C=N_NUCLEONS) and N(C=N_COLLISONS) are the number of participating nucleons and binary collisions. The statistical errors are negligible and only systematic errors are quoted. COL(NAME=CENTRALITY) is centrality.

Dilepton mass spectra in p+p collisions at s(1/2) = 200-GeV and the contribution from open charm**

The PHENIX collaboration Adare, A. ; Afanasiev, S. ; Aidala, C. ; *et al.*

Phys.Lett.B 670 (2009) 313-320, 2009.

Inspire Record 778611 DOI 10.17182/hepdata.73669

The phenix experiement has measured the electron-positron pair mass spectrum from 0 to 8 GeV/c^2 in p+p collisions at sqrt(s)=200 GeV. The contributions from light meson decays to e^+e^- pairs have been determined based on measurements of hadron production cross sections by phenix. They accou is dominated by semileptonic decays of charmed hadrons correlated through flavor conservation. Using the spectral shape predicted by PYTHIA, we estimate the charm production cross section to be 544 +/- 39(stat) +/- 142(syst) +/- 200(model) \mu b, which is consistent with QCD calculations and measure

2 data tables match query

Table 1 Differential charm cross section at mid rapidity An additional +-39.5 microbarn error, due to the validity of the model used to extrapolate the data, is not included The contribution from beauty estimated to be 3.7 microbarn, has been subtracted. The c->e branching ratio u

Table 2 Total charm cross section An additional systemactic error of +- 200 microbarn, due to the validity of the model used to extrapolate the data, is not included. To obtain the total charm cross section, the differential charm cross section has been extrapolated to the whole rap

Measurement of the mid-rapidity transverse energy distribution from s(NN)(1/2) = 130-GeV Au + Au collisions at RHIC**

The PHENIX collaboration Adcox, K. ; Adler, S.S. ; Ajitanand, N.N. ; *et al.*

Phys.Rev.Lett. 87 (2001) 052301, 2001.

Inspire Record 555603 DOI 10.17182/hepdata.31419

The first measurement of energy produced transverse to the beam direction at RHIC is presented. The mid-rapidity transverse energy density per participating nucleon rises steadily with the number of participants, closely paralleling the rise in charged-particle density, such that E_T / N_ch remains relativ at sqrt(s_NN)=130 GeV is at least epsilon_Bj = 4.6 GeV/fm^3 which is a factor of 1.6 larger than found at sqrt(s_NN)=17.2 GeV (Pb+Pb at CERN).

1 data table match query

Table 1 130 GeV is sqrt(S) per nucleon-nucleon collision. The statistical errors are negligible and only systematic errors are quoted. COL(NAME=CENTRALITY) is centrality.

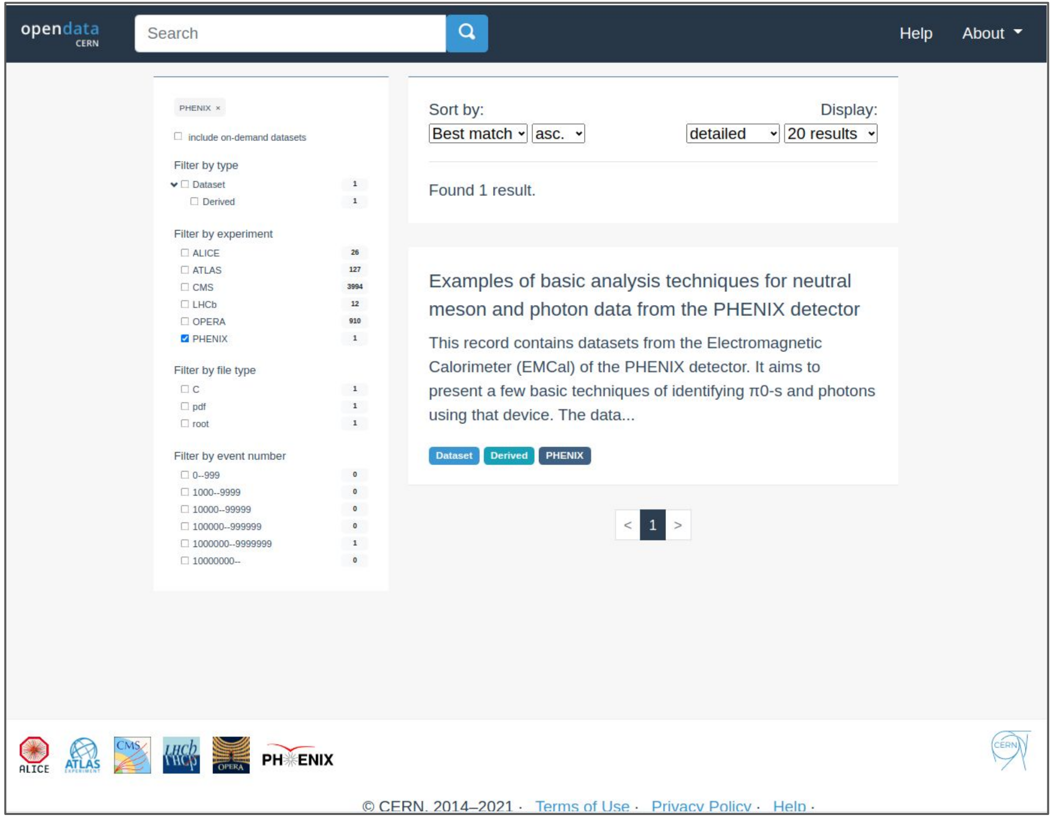
HEPData submissions mandated for all new publications

Revisiting older publication materials as time permits

Using GitHub for material development, support of team effort

Your contributions are welcome!

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



PHENIX ×

☐ include on-demand datasets

Filter by type

☐ Dataset	1
☐ Derived	1

Filter by experiment

☐ ALICE	26
☐ ATLAS	127
☐ CMS	3994
☐ LHCb	12
☐ OPERA	910
☑ PHENIX	1

Filter by file type

☐ C	1
☐ pdf	1
☐ root	1

Filter by event number

☐ 0--999	0
☐ 1000--9999	0
☐ 10000--99999	0
☐ 100000--999999	0
☐ 1000000--9999999	1
☐ 10000000--	0

Sort by: Best match, asc.

Display: detailed, 20 results

Found 1 result.

Examples of basic analysis techniques for neutral meson and photon data from the PHENIX detector

This record contains datasets from the Electromagnetic Calorimeter (EMCal) of the PHENIX detector. It aims to present a few basic techniques of identifying π0-s and photons using that device. The data...

Dataset Derived PHENIX

< 1 >

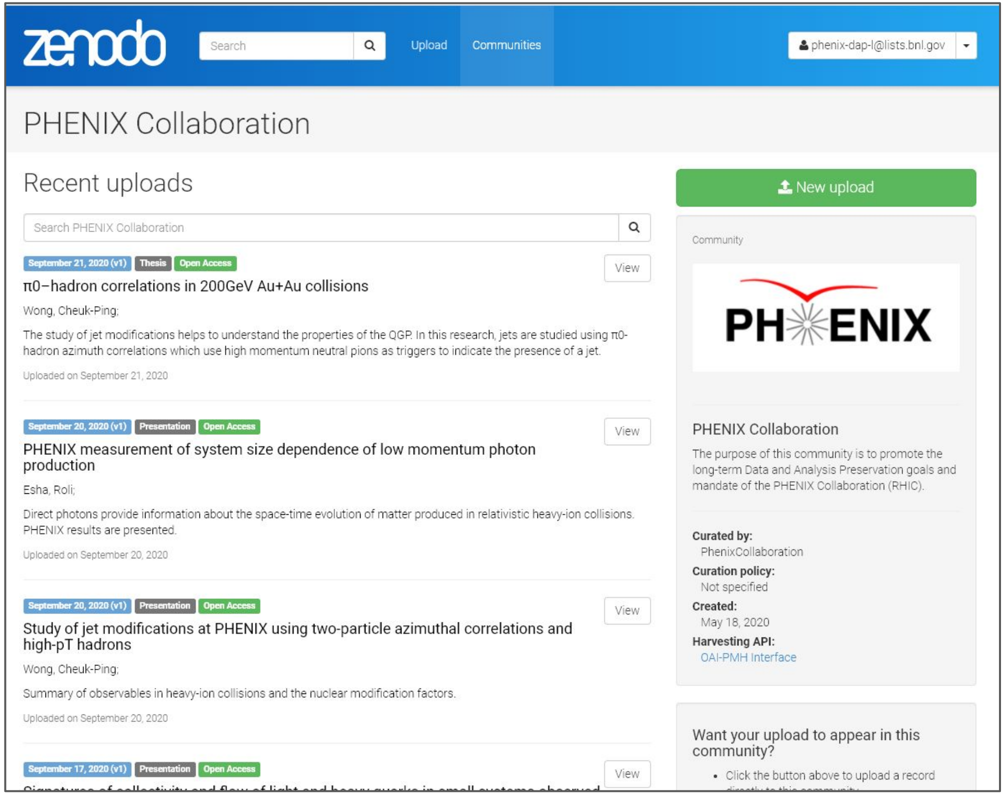
© CERN, 2014–2021 · Terms of Use · Privacy Policy · Help ·

- Package Content:
  - 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



PHENIX Collaboration

Recent uploads

September 21, 2020 (v1) Thesis Open Access

π0−hadron correlations in 200GeV Au+Au collisions

Wong, Cheuk-Ping;

The study of jet modifications helps to understand the properties of the QGP. In this research, jets are studied using π0-hadron azimuth correlations which use high momentum neutral pions as triggers to indicate the presence of a jet.

Uploaded on September 21, 2020

September 20, 2020 (v1) Presentation Open Access

PHENIX measurement of system size dependence of low momentum photon production

Esha, Roli;

Direct photons provide information about the space-time evolution of matter produced in relativistic heavy-ion collisions. PHENIX results are presented.

Uploaded on September 20, 2020

September 20, 2020 (v1) Presentation Open Access

Study of jet modifications at PHENIX using two-particle azimuthal correlations and high-pT hadrons

Wong, Cheuk-Ping;

Summary of observables in heavy-ion collisions and the nuclear modification factors.

Uploaded on September 20, 2020

September 17, 2020 (v1) Presentation Open Access

Signatures of collectivity and flow of light and heavy quarks in small systems observed

PHENIX Collaboration

The purpose of this community is to promote the long-term Data and Analysis Preservation goals and mandate of the PHENIX Collaboration (RHIC).

Curated by: PhenixCollaboration
Curation policy: Not specified
Created: May 18, 2020
Harvesting API: OAI-PMH Interface

Want your upload to appear in this community?

- Click the button above to upload a record

Docker/REANA: more in the interactive section

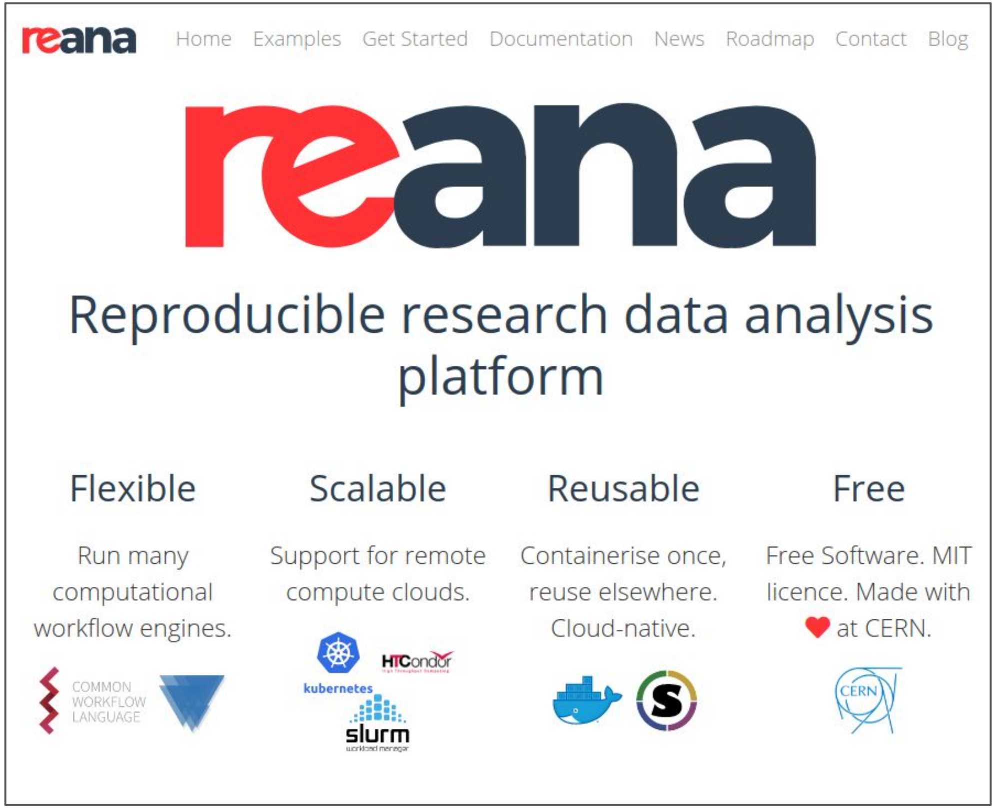
...Just a quick intro here...

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 (e.g. fun4all etc)
- Created images to preserve legacy ROOT5 versions to ensure compatibility with older macros
    - You can have an environment on your laptop identical to an interactive SDCC node i.e. same version of ROOT, emacs, X11
- 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 - reproducible analysis

- **Deployed at BNL**, with PHENIX team currently on the learning curve, running analysis macros
- Demonstrated capability to run complex analyses at the LHC scale
- 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
- We will have an interactive exercise in a few minutes



Reproducible research data analysis platform

Flexible

Run many computational workflow engines.

Scalable

Support for remote compute clouds.

Reusable

Containerise once, reuse elsewhere. Cloud-native.

Free

Free Software. MIT licence. Made with ❤ at CERN.

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

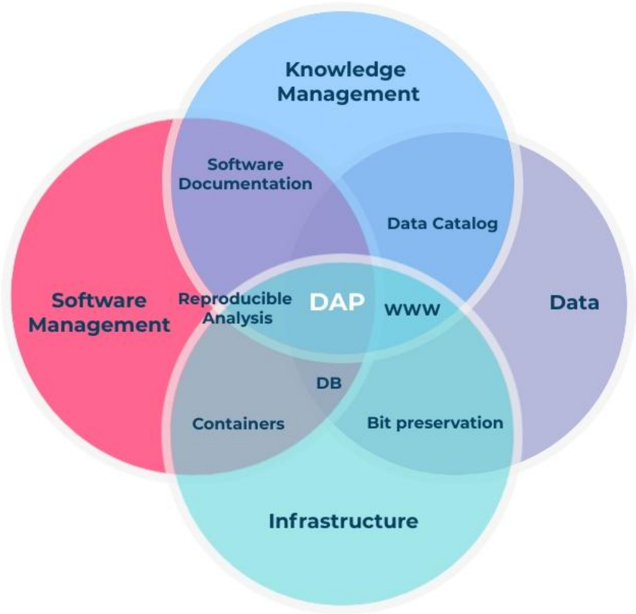
- In the past, there was an assumption that while DAP required investment and effort upfront, any benefits coming from it could only be realized in the long-term, thus complicating adoption of DAP practices. This may no longer be true.
- The technology landscape has changed.
- Many aspects of DAP overlap 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 collaboratior 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 in new experiments: *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)
- Prioritize analyses for preservation as effort is always limited

Final thoughts

- In the past two years PHENIX made progress in Analysis Preservation
- A new website has been commissioned, designed for low-maintenance, long-term operation. It serves as a portal to a external resources leveraged by PHENIX
  - HEPData, Zenodo, GitHub, Docker Hub, InspireHEP, OpenData
- We are catching up with other experiments in the area of HEPData submissions
- 400 uploads on Zenodo - an impressive body of work, all theses, conference presentations for the past few years, older technical papers
- We created our first entry on the CERN OpenData portal and will continue this effort
- Wider adoption of REANA is a worthy goal
  - If your analysis depends on “pure ROOT” macros it’s a low hanging fruit
- Creation of full PHENIX containers is on the critical path and is in progress