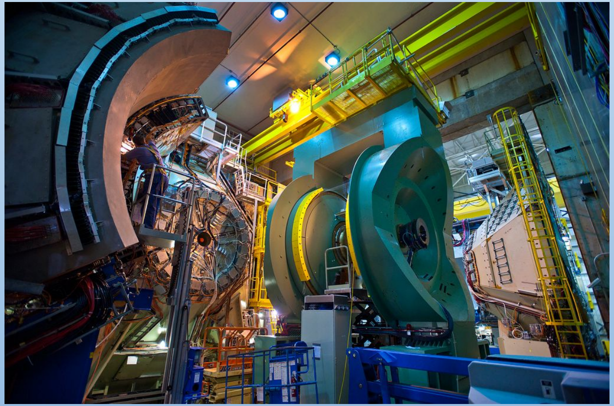
Data & Analysis Preservation: Experience in PHENIX

Maxim Potekhin
Nuclear and Particle Physics Software Group

Brookhaven National Laboratory



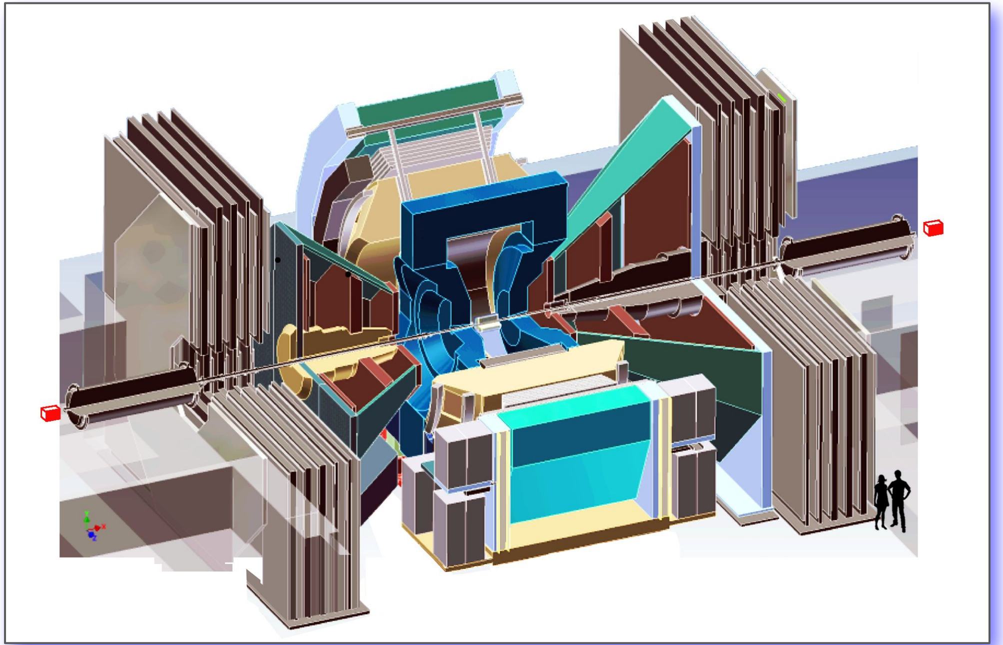
RHIC & AGS Users Meeting 2021

06/10/2021

PHENIX

PHENIX

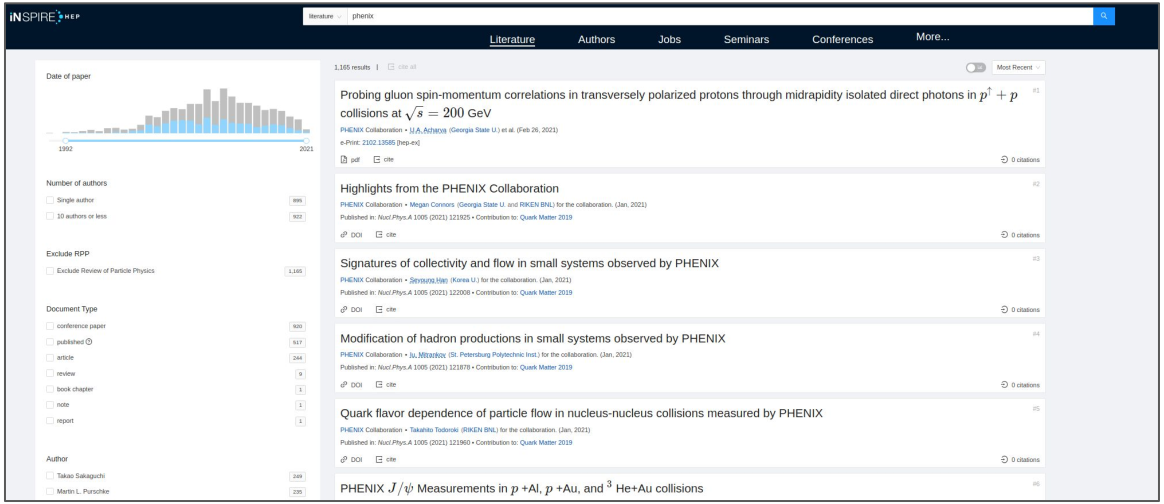
- "Pioneering High Energy Nuclear Interaction eXperiment"
- One of the two large RHIC experiments
- A large, complex general purpose detector with a considerable physics reach and *complex analyses*



- Please see the "PHENIX Collaboration Community" on Zenodo, the CERN-based digital repository: https://zenodo.org/communities/phenixcollaboration/

PHENIX today

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



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.

Data and Analysis Preservation in 2020s

- In the past decade, 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.
- There are challenges in preservation of the necessary software and infrastructure elements in an organized and functional state.
  - However, even if the software is well organized and documented it won't be useful without detailed knowledge of how to apply it in the specific domain
  - DAP thus brings focus to knowledge management - which is also conducive to the quality of the current research i.e. has short- and medium-term impact
- DAP is aligned with the goals of "software sustainability" (a relatively new term)
- In this presentation, we are sharing the experience, technology choices and status of the effort of the PHENIX Collaboration in this area.

DAP: quotes from the experts

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

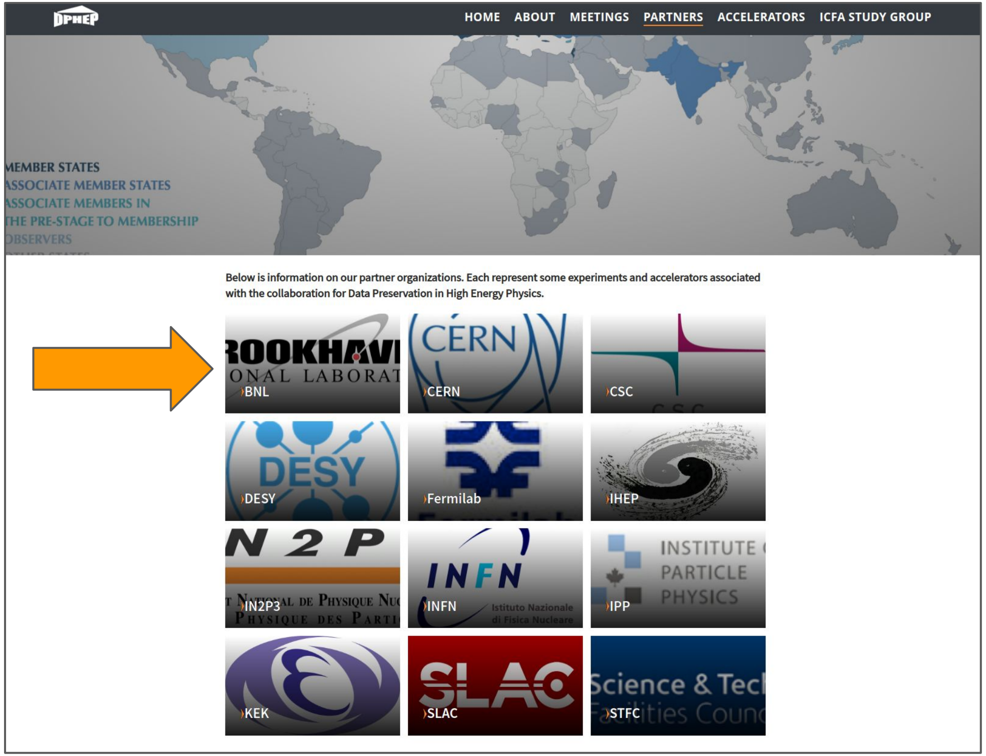
- DAP universally depends on continuity of services and expertise provided by the facility.
- This is especially true for PHENIX: BNL SDCC is its only functioning computing site.
- In addition to bit preservation (mass storage) the facility provides software builds and provisioning capabilities (including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

Nuclear and Particle Physics Software Group

- https://npps.bnl.gov/
- *“The Nuclear and Particle Physics Software (NPPS) Group in Brookhaven National Laboratory's Physics Department participates in a wide range of experiments across BNL's nuclear and particle physics programs. NPPS provides software and expertise across many technical areas, with a particular emphasis on common software solutions.”*
- NPPS provides DAP support for PHENIX and is a natural venue for cross-experiment and cross-site development and collaboration in this area.

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

International Collaboration for Data Preservation and Long Term Analysis 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.”

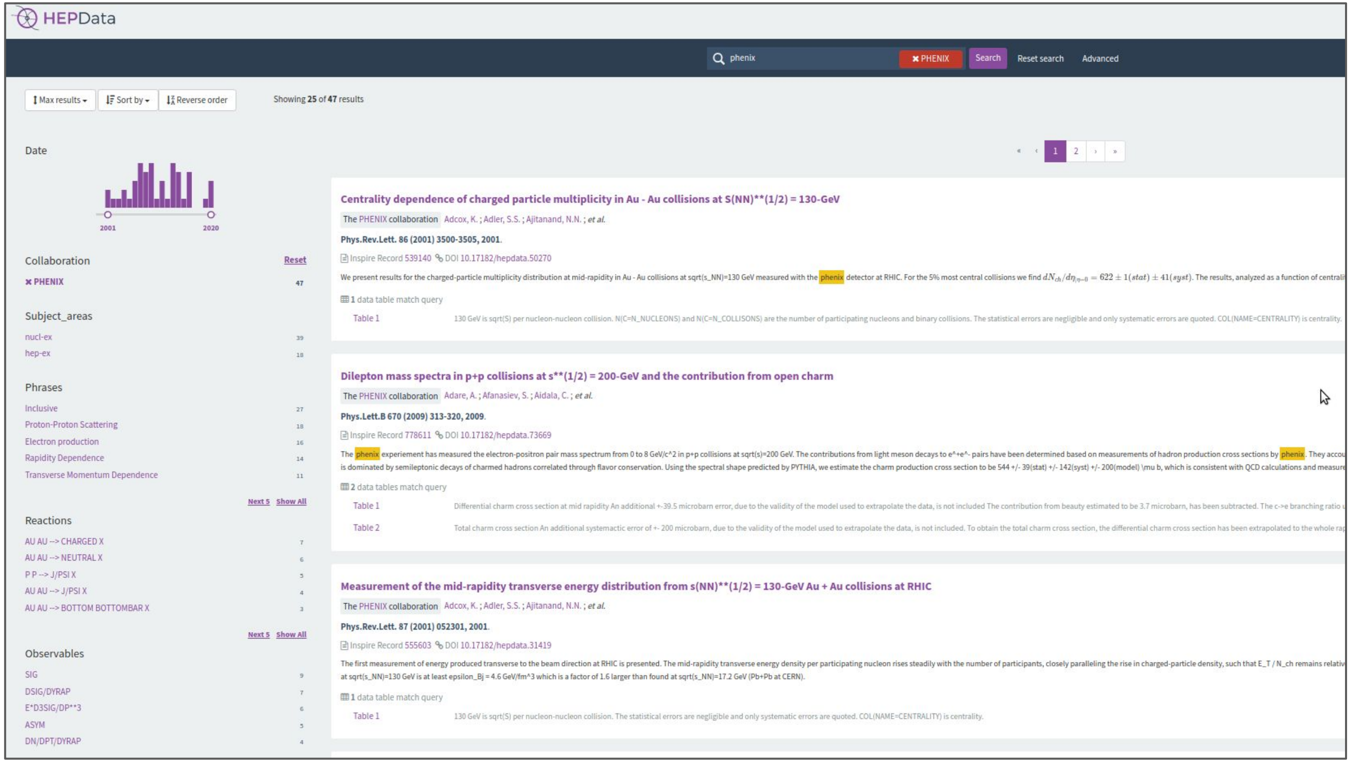
- SDCC (BNL) is an active DPHEP partner on the facility side, participating in DAP technology development and testing. This gives BNL and the RHIC community optimal access to the state-of-the-art methodologies and tools.

- PHENIX members participated in the DPHEP Workshop at CERN in 2019 and continue to be actively engaged with and receive guidance from DPHEP

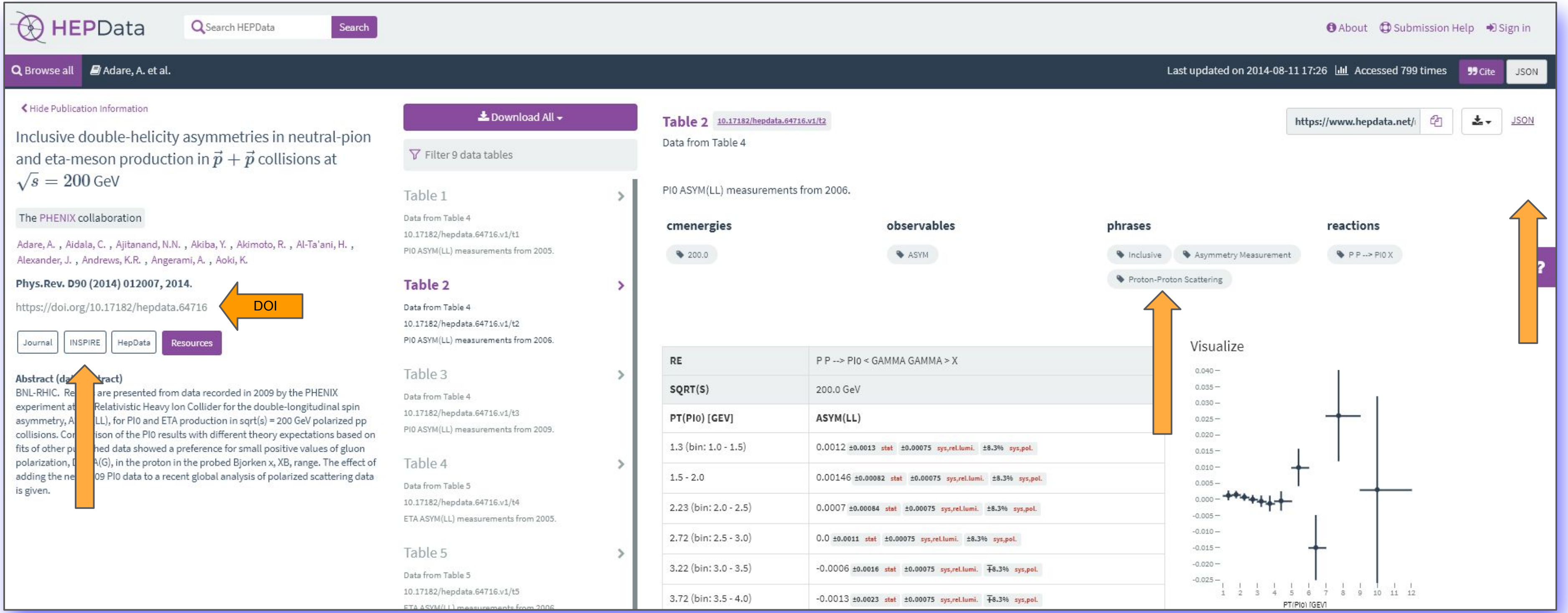
DPHEP: definition of standard tiers of Data Preservation

- 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.

HEPData - a portal for data used in publications



HEPData - useful features

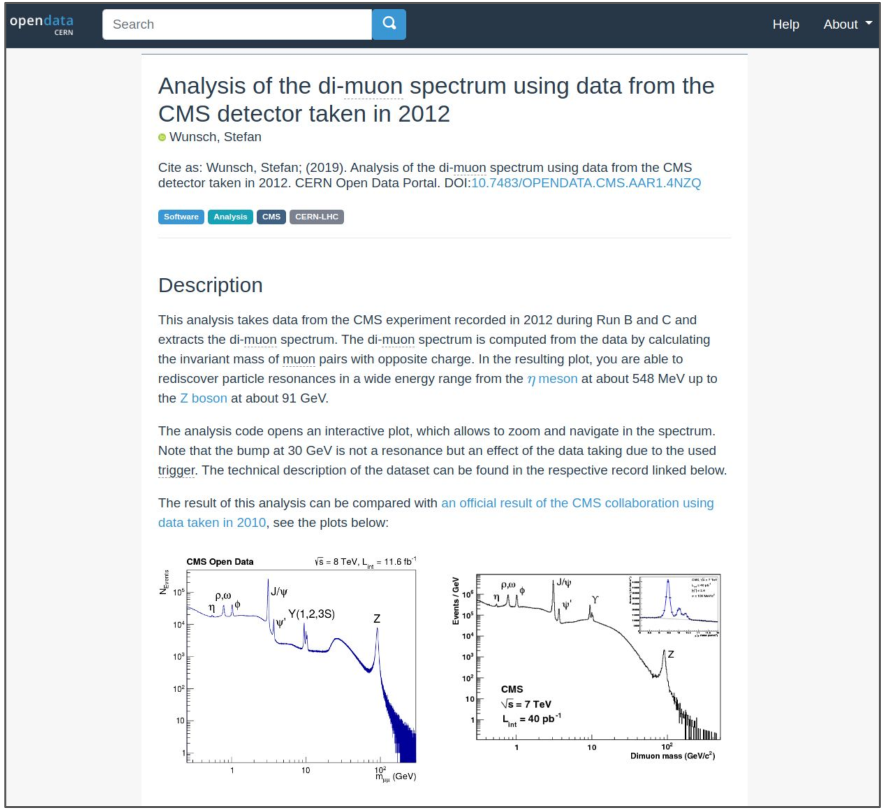


HEPData as a part of the PHENIX DAP work

- hepdata.net
- Durable data storage and discovery with minted DOIs, for preservation of numerical data used in publications e.g. plots and tables
    - To be uploaded, the data needs to be formatted according to HEPData specifications (YAML)
    - Formatting and validation can require substantial work, depending on the material
- Data can be easily exported in a standard format to facilitate comparisons with other experiments and theory
- Currently PHENIX has committed 47 entries to this portal vs its total of 240 publications
    - PHENIX publication policy has been updated in 2020 to mandate HEPData submission packages for all new publications, and work on older items happens “as time permits”
    - We use GitHub to develop and review materials, the process is well understood and well organized - however lack of available effort remains a problem
    - Thanks to C.Nattrass for many contributions to this work area

OpenData

- OpenData - a versatile system for aggregating and preserving **software, data and documentation** pertaining to analyses published by the experiments
- PHENIX is finalizing its first contribution to this portal, containing NTuples and ROOT macros which illustrate analyses of the data with the Electromagnetic Calorimeter
- NB. The system mints DOIs



PHENIX EMCal - an OpenData entry (work in progress)

system ($x > 0$ is the West Arm, negative x is South).

The *MBntup.root* file is produced from minimum bias data (no lower limit on single cluster p_T in *gnt* or pair p_T in *ggntuple*), whereas in *ERTntup.root* the threshold for single cl... p_T in *gnt* is 5 GeV, and the threshold for pair p_T in *ggntuple* is also 5 GeV. Note that ... here we restrict only the pair p_T, the energy of the individual clusters can be (and ofte... significantly lower.

Variable name	Description
cent	Event centrality
vtxZ	z-vertex of the event
pt	Transverse momentum of t...
costheta	Polar angle of the cluster (...
phi	Azimuthal angle of the clus...
sec	EMCal sector of the cluster (γ-candidate)
ecore	"Core" energy of the cluster (γ-candidate)
ecent	Energy in the central tower of the cluster (γ-candidate)
tof	Time-of-flight in the central tower of the cluster (γ-candidate)
prob	Probability that the cluster is a photon (based on χ^2)
disp	Dispersion of the cluster (γ-candidate)
chisq	χ^2 from expected photon shape of the cluster (γ-candidate)
twrhit	Number of towers in the cluster (γ-candidate)
stoch	Combined variable to describe "photonness" of the cluster (γ-candidate)
x	x-position of impact point on the EMCal surface
y	y-position of impact point on the EMCal surface
z	z-position of impact point on the EMCal surface

```
ggntuple->Draw("mass","mass<1.0");
ggntuple->Draw("mass","mass<0.4&&pt>8.0");
ggntuple->Draw("mass>>htemp1","mass<0.4");
ggntuple->Draw("mass>>htemp2","mass<0.4&&chisq1<2.0&&chisq2<2.0");
htemp1->SetLineColor(1);
...(2);
```

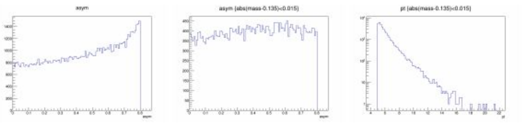


FIG. 2. ERT data, plots from the pair ntuple. Left: energy asymmetry distribution for all pairs.

... see Fig. 1.

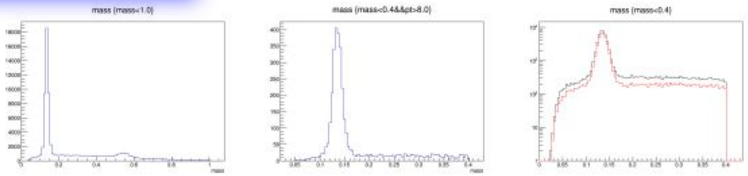
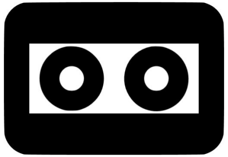


FIG. 1. ERT data, plots from the pair ntuple. Left: Invariant mass in the 0-1 GeV region. You can see a strong π^0 and a well-recognizeable η peak. Middle: π^0 peak for pairs with p_T greater than 8 GeV/c. You can clearly see the combinatorial background outside the peak, which should

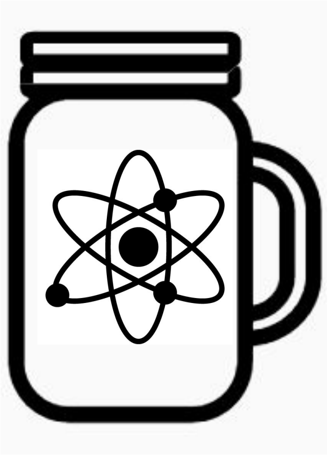
PHENIX: Challenges of **K**nowledge **M**anagement

- Need to keep records of software provenance, dependencies, configuration, user pattern (e.g. which specific macros/libraries run, arguments, sequence, conditions...)
- "Data artifacts" such as conditions- and calibrations-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
  - For example, analysis-specific dead channel maps, recalculated efficiencies, "good run" lists etc
  - Fiducial cuts specific to the analysis (not always documented for reuse)
  - Numerical data in the code (unclear provenance)
- Information spread across a few legacy web resources - for the software, detector and subsystem information and other documentation
- There is a required "KM" section in the Analysis Note template, but its efficacy is not always optimal
- Lots of "moving" parts, not easy to capture and document
  - Even more difficult after the analysis is done, paper published and the team moves on

The DAP Strategy in PHENIX

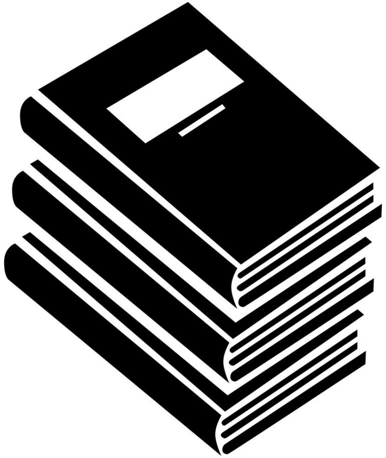


Bit preservation



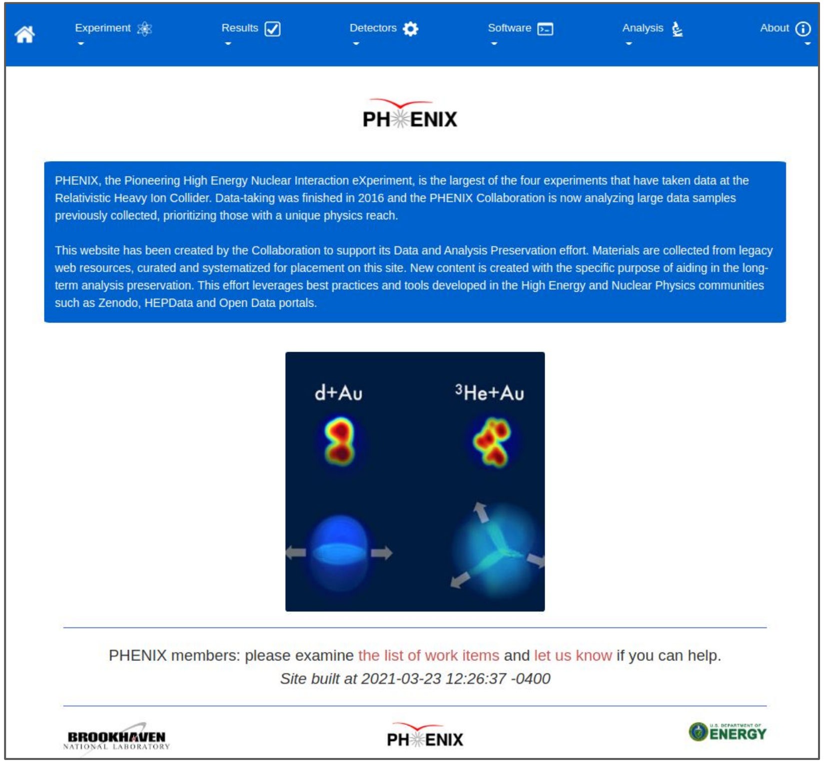
Analysis capture

Improved
Web-based documentation



A highly-functional repository for research materials

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



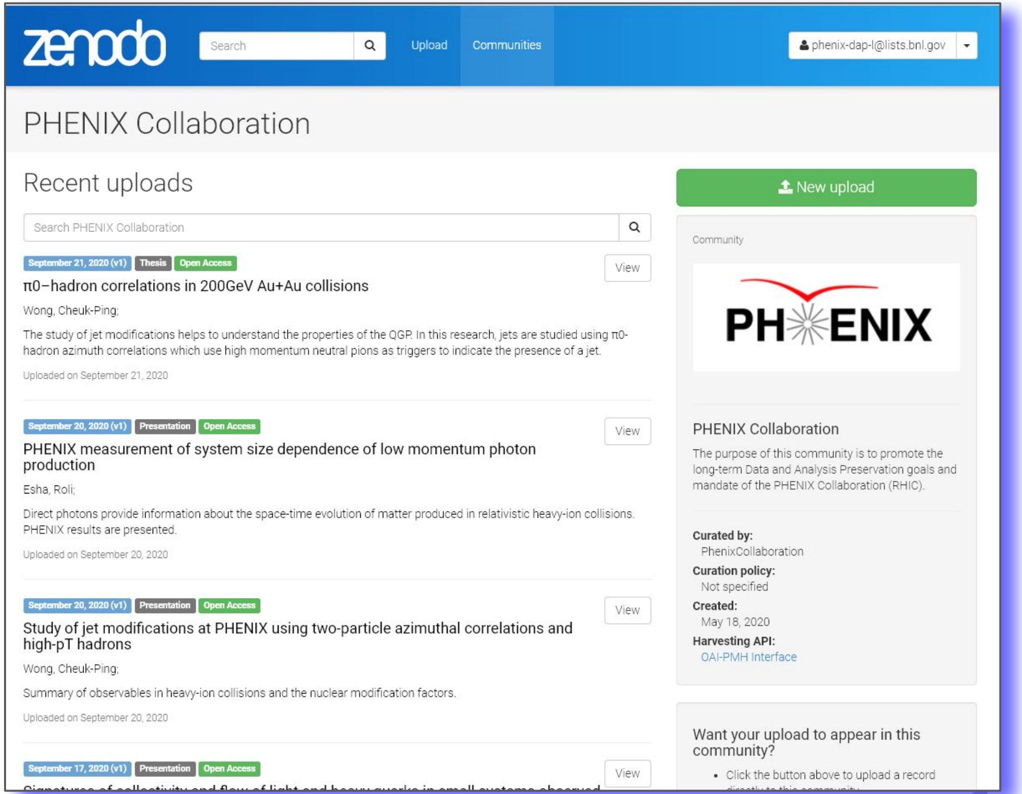
The website as the main portal to PHENIX info

- Links to the many resources are provided and managed on the new PHENIX Website:
    - Zenodo (a digital repository developed and maintained by CERN)
    - HEPData, OpenData
    - InspireHEP
    - GitHub, Docker Hub
    - REANA
    - Technical notes and descriptions of the detector subsystems, run history etc
- The website effectively replaced legacy web resources which were getting more and more difficult to maintain
- Design philosophy - ease of long term maintenance and security
- We are using a static site generator (Jekyll) to achieve these goals

Zenodo@CERN - the PHENIX community

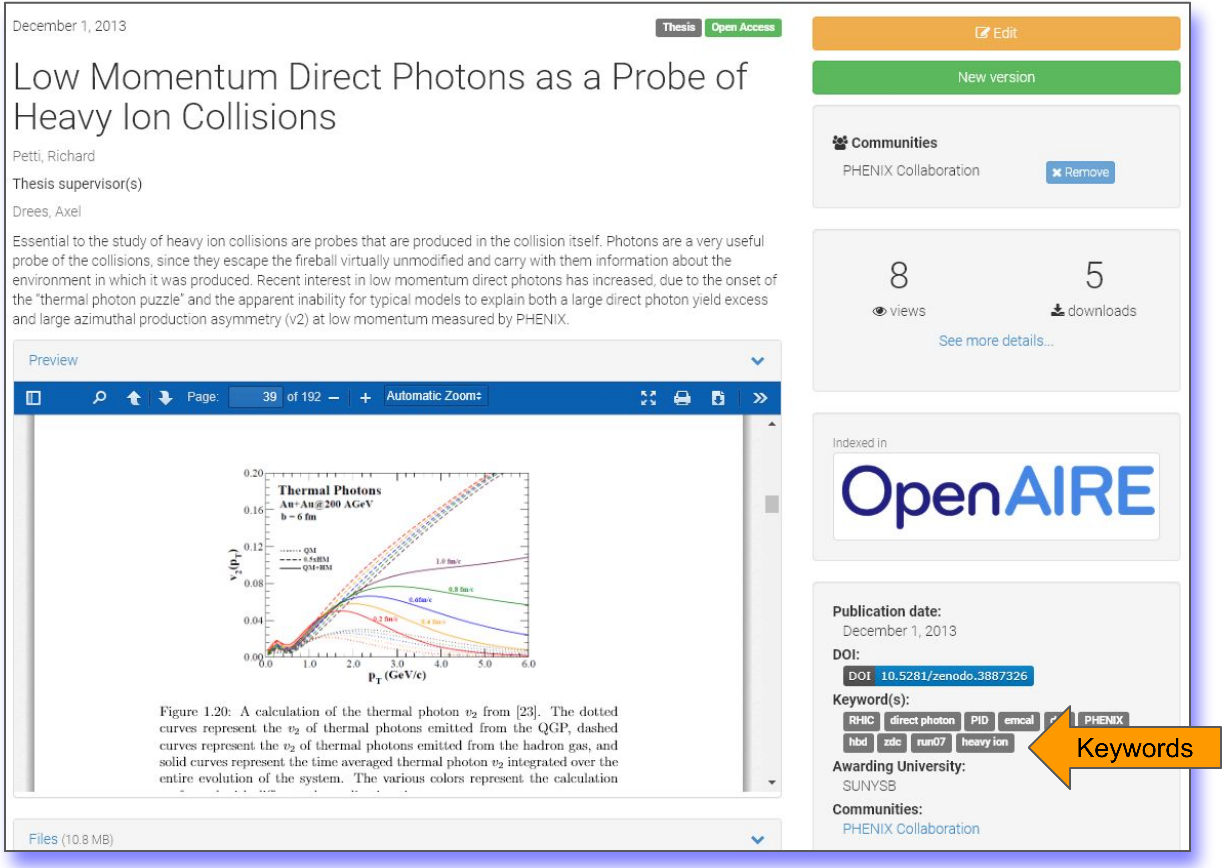
https://zenodo.org/communities/phenixcollaboration

- Zenodo is a world-class digital repository
- PHENIX-branded
- Curated
- Discoverable
- Well-suited for long-term preservation, with **DOIs**
- Carefully indexed (keywords are managed on the PHENIX website)
- Elastic search capability



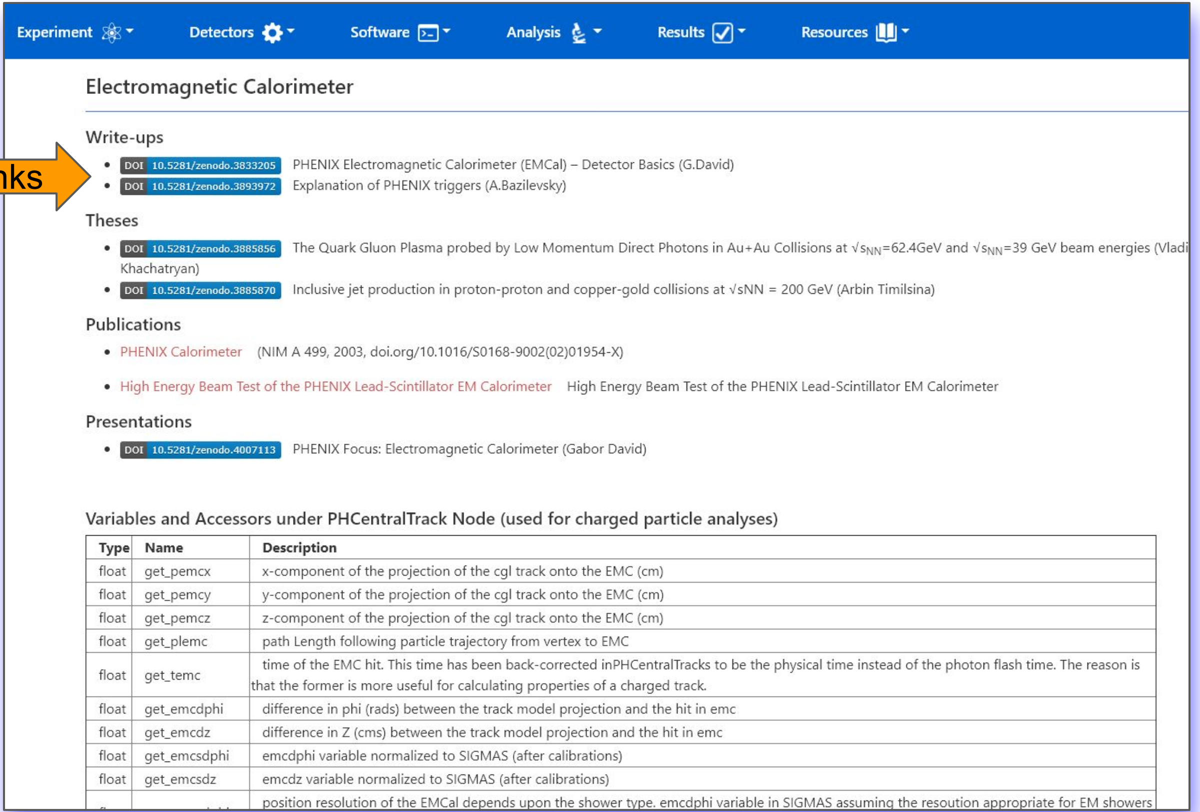
An example of a PHENIX item on Zenodo

- PhD theses are committed to Zenodo and tagged with keywords
- Conference contributions for the past few years are committed to Zenodo



An example of a subsystem page on the new website

Zenodo links



Experiment | Detectors | Software | Analysis | Results | Resources

Electromagnetic Calorimeter

Write-ups

- DOI 10.5281/zenodo.3833205 PHENIX Electromagnetic Calorimeter (EMCal) – Detector Basics (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 (Vladi 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)

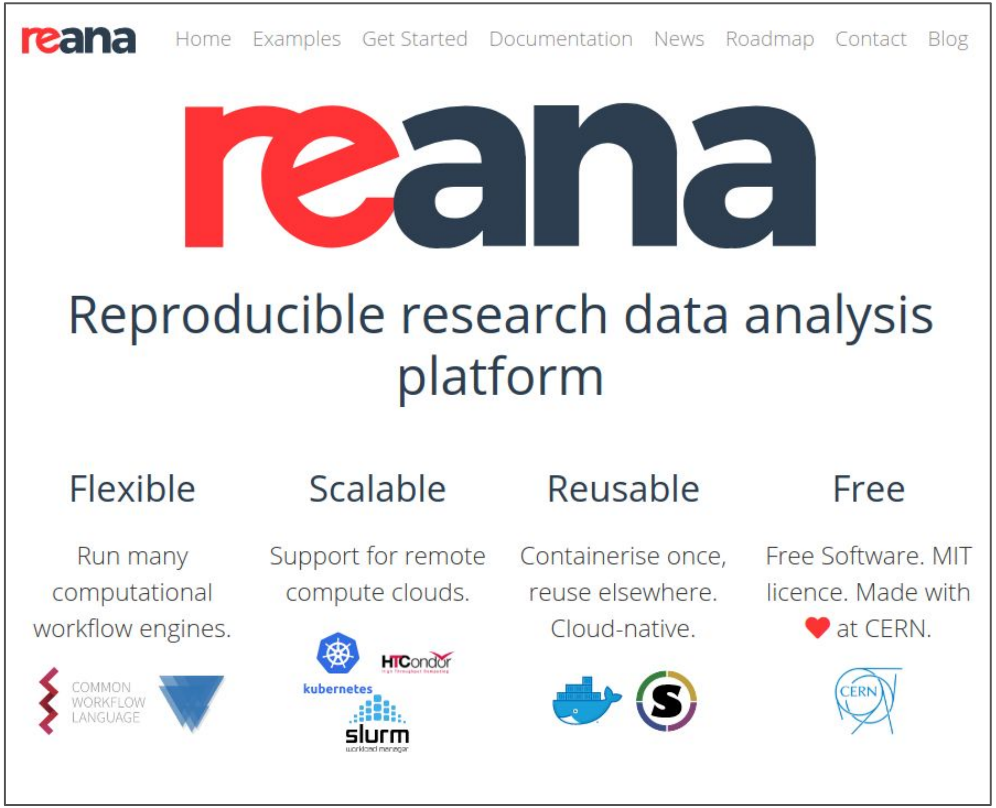
Type	Name	Description
float	get_pemcx	x-component of the projection of the cgl track onto the EMC (cm)
float	get_pemcy	y-component of the projection of the cgl track onto the EMC (cm)
float	get_pemcz	z-component of the projection of the cgl track onto the EMC (cm)
float	get_plemc	path Length following particle trajectory from vertex to EMC
float	get_temc	time of the EMC hit. This time has been back-corrected inPHCentralTracks to be the physical time instead of the photon flash time. The reason is that the former is more useful for calculating properties of a charged track.
float	get_emcdphi	difference in phi (rads) between the track model projection and the hit in emc
float	get_emcdz	difference in Z (cms) between the track model projection and the hit in emc
float	get_emcsdphi	emcdphi variable normalized to SIGMAS (after calibrations)
float	get_emcsdz	emcdz variable normalized to SIGMAS (after calibrations)
		position resolution of the EMCal depends upon the shower type. emcdphi variable in SIGMAS assuming the resoution appropriate for EM showers

Capturing the Software Environment

- Most of the PHENIX software is not portable e.g. cannot be built and used on an arbitrary system outside of SDCC
- 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 on CVMFS
  - Creating a custom image by selecting most relevant libraries
- 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: https://reanahub.io/

- Demonstrated in real-life, large scale analysis scenarios at CERN
- The software environment is captured in containers
- The workflow is captured by using a structured description (YAML) with options for both linear workflows and arbitrary DAGs
- Increasingly popular in projects including the EIC
- **Deployed at BNL**, with PHENIX team currently on the learning curve, running analysis macros
- Both storage and CPU can be scaled up if resources available
- Tutorials are being prepared for the PHENIX School’21



Lessons learned

- DAP: *plan and start early* (should be a part of someone's job description)
  - The effort will pay for itself by increasing productivity
  - PHENIX is fighting an uphill battle here due to a late start
- Avoid building in-house information systems, there are plenty of 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
- A lot of potential for DAP collaboration across many projects
  - Including future experiments e.g. EIC