Data & Analysis Preservation in PHENIX: status of the website and future plans

Maxim Potekhin
Nuclear and Particle Physics Software Group

BROOKHAVEN NATIONAL LABORATORY

PHENIX EC Meeting
01/14/2021

PHENIX

This presentation

- An introduction to the original version of the PHENIX redesigned website was presented to the EC in March 2020:
  - https://docs.google.com/presentation/d/1Vnq0wl9PyZKYmfRw_k9ZOHko6IwZDDQymUixDA-Srvg/edit?usp=sharing
  - Please see the link above for a more generic discussion of motivations, technology choices etc
- Today we shall review the current status of the website and how it leverages the available cloud resources
- ...and discuss ways to ensure usefulness of the website as a component of Analysis Preservation and its potential to advance more immediate goals of PHENIX
- Some auxiliary info can be found in the “backup slides”

Some background

- The legacy PHENIX website had been showing its age for a while
  - Some useful information was spread over multiple web servers
  - Used as a document repository with limited functionality and lack of external visibility
  - Certain software modules in its framework were potentially vulnerable security-wise
- Technical problems encountered in 2020 illustrate operational risks due to difficulty of upgrades e.g. maintaining the legacy software stack in the evolving OS environment while staying abreast of security requirements
- Less DB support available as time goes on (actually none at this point)

Strategy 2020

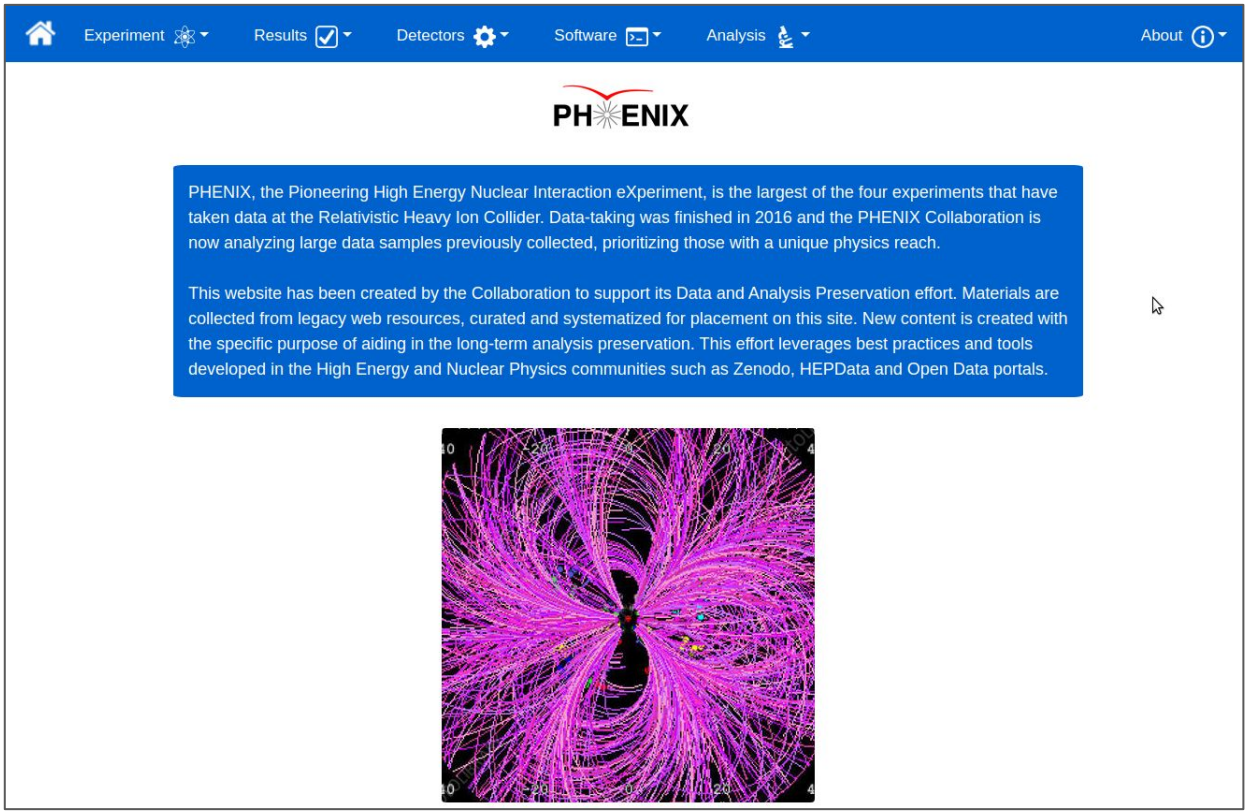
- DAP = Data and Analysis Preservation
- Design a new website aligned with the goals of DAP, with emphasis on ease of maintenance and long-term durability
- Collect and curate materias from available resources, create new ones where possible and necessary
- Factor out the repository and other similar functionality by leveraging modern solutions developed and employed in the NP and HEP communities
    - The website effectively becomes a portal to cloud resources, with a modest amount of material still hosted locally

The website technology

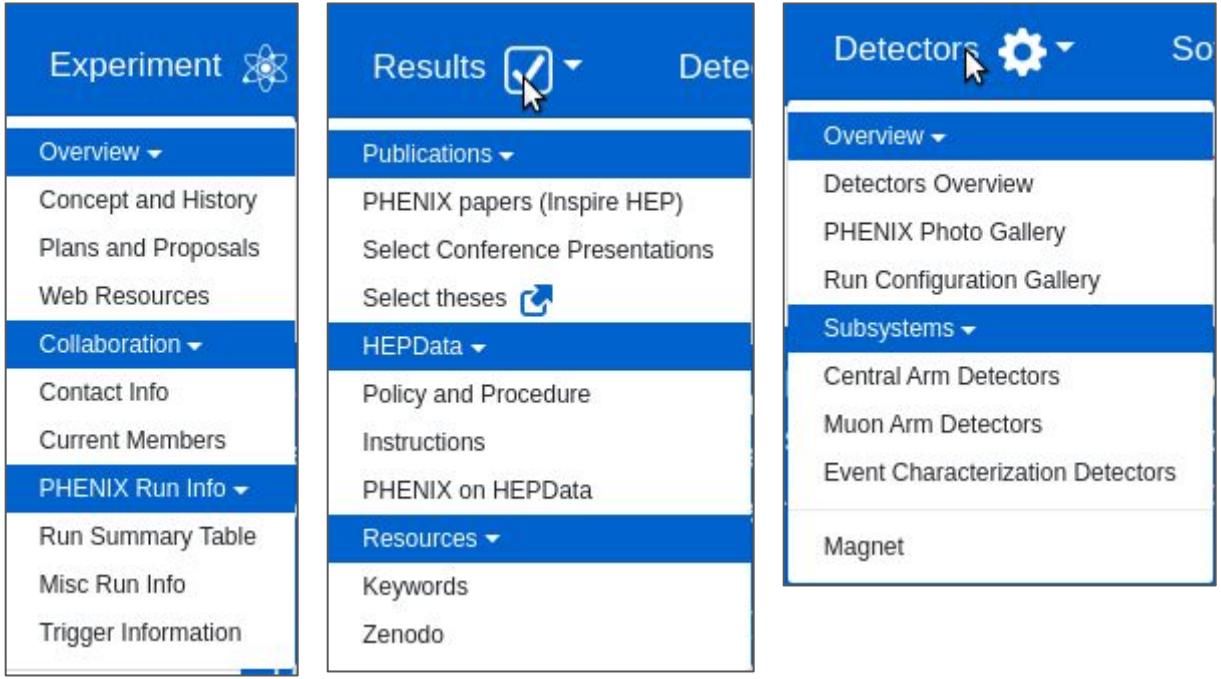
- Use a static website generator for speed and security
- Content and layout are separate; easy to use markup (not HTML)
- Data kept in specially formatted files (YAML) to provide DB-like functionality without having an actual database, also flexibility and consistency across multiple pages - define data once, use in many places
- Free hosting of the development version on GitHub pages
- As of Fall 2020 the production version is officially hosted at BNL at the canonical URL: https://www.phenix.bnl.gov/
- In the following few slides we will review the site

Website: the landing page

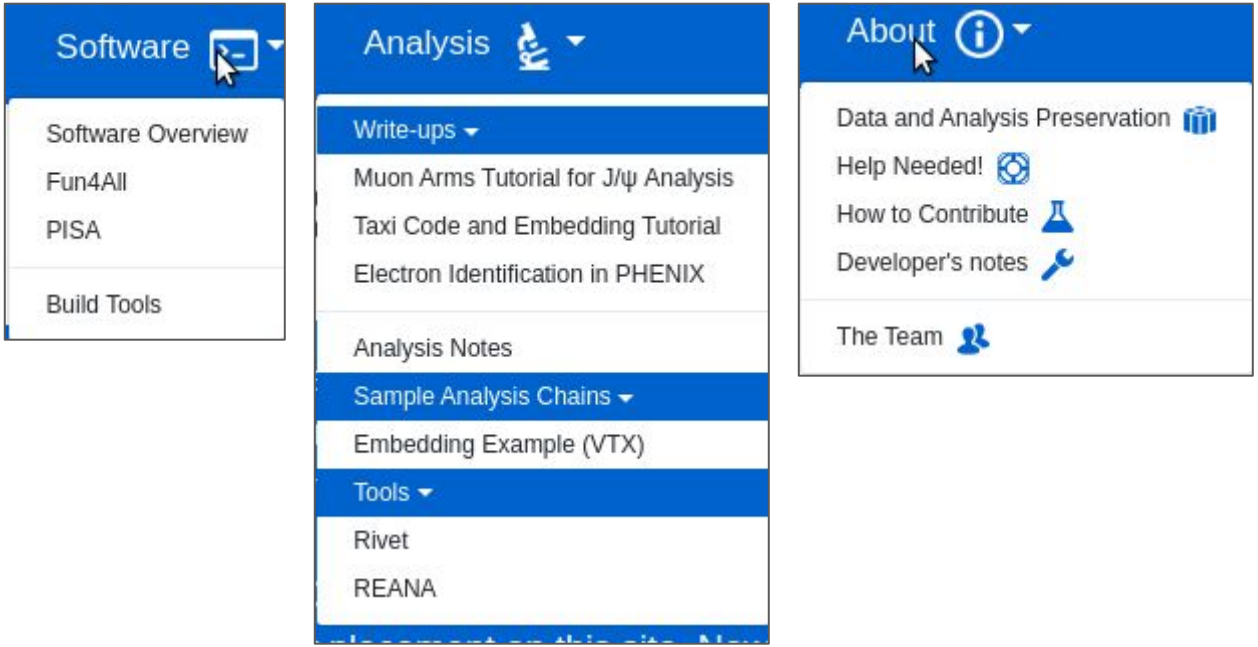
Menus



Website: dropdown menus (1)



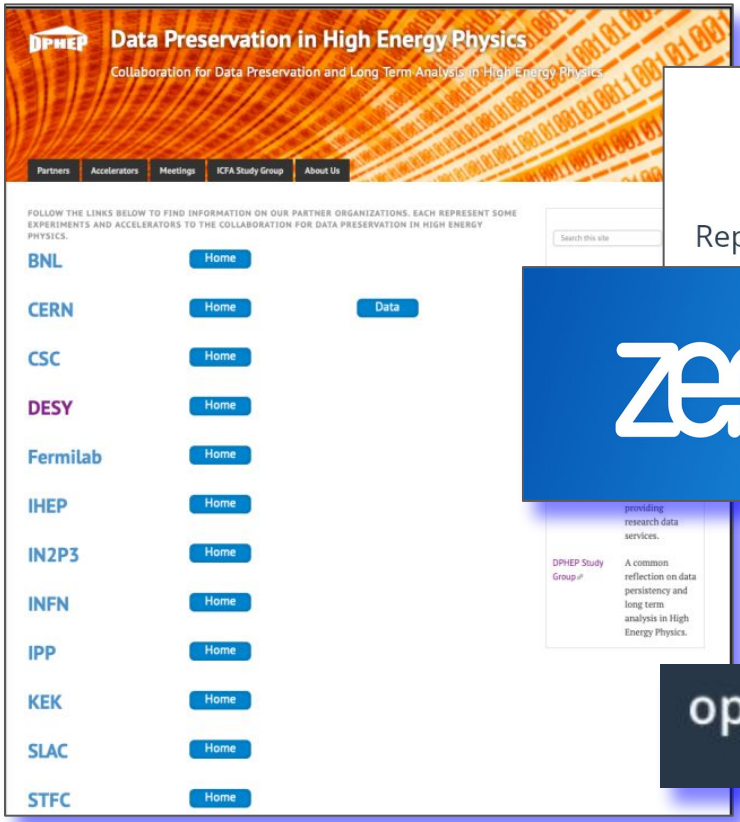
Website: dropdown menus (2)



Leveraging the community cloud resources

- Use well established solutions widely adopted in the community
  - NB. solid back-end and support
  - Look at best practices at CERN and other labs
  - Reduces cost and increases long-term viability of the PHENIX Web infrastructure
- Resources already adopted by PHENIX
  - Zenodo (a digital repository)
  - HEPData - a repository for numerical data used in publications
  - InspireHEP - a searchable publication catalog
- i.e. this type of functionality effectively migrated to the cloud from the legacy site
- Heavy investment in Zenodo and HEPData
- Upcoming: OpenData and REANA

CERN and community tools for DAP



reana

Reproducible research data analysis platform

HEPData

zenodo

iNSPIRE HEP

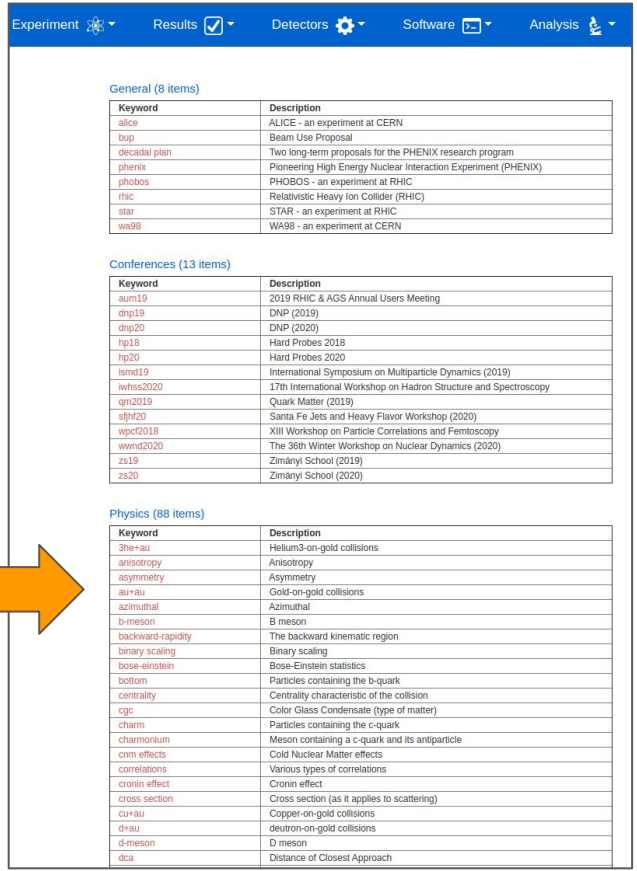
opendata CERN

Zenodo

- Zenodo - close to 200 PHENIX items made available on the portal
- Let me point out once again that DOIs are an important feature for long-term preservation
- Uploaded/tagged ⅔ of PHENIX theses (currently at 124), work in progress with completion date in 2021 - thanks to Stacyann Nelson
- Steady upload of conference/workshop and other presentation, with a dedicated page on the new site
- Keywords - essential for discoverability and integrity of references
    - A list is carefully curated and used with all uploads

Zenodo keywords page

Working query links



Experiment | Results | Detectors | Software | Analysis

General (8 items)

Keyword	Description
alice	ALICE - an experiment at CERN
bup	Beam Use Proposal
decadal plan	Two long-term proposals for the PHENIX research program
phenix	Pioneering High Energy Nuclear Interaction Experiment (PHENIX)
phobos	PHOBOS - an experiment at RHIC
rhic	Relativistic Heavy Ion Collider (RHIC)
star	STAR - an experiment at RHIC
wa98	WA98 - an experiment at CERN

Conferences (13 items)

Keyword	Description
aum19	2019 RHIC & AGS Annual Users Meeting
dnp19	DNP (2019)
dnp20	DNP (2020)
hp18	Hard Probes 2018
hp20	Hard Probes 2020
ismd19	International Symposium on Multiparticle Dynamics (2019)
iwhss2020	17th International Workshop on Hadron Structure and Spectroscopy
qm2019	Quark Matter (2019)
sfjhf20	Santa Fe Jets and Heavy Flavor Workshop (2020)
wpcf2018	XIII Workshop on Particle Correlations and Femtoscopy
wwnd2020	The 36th Winter Workshop on Nuclear Dynamics (2020)
zs19	Zimányi School (2019)
zs20	Zimányi School (2020)

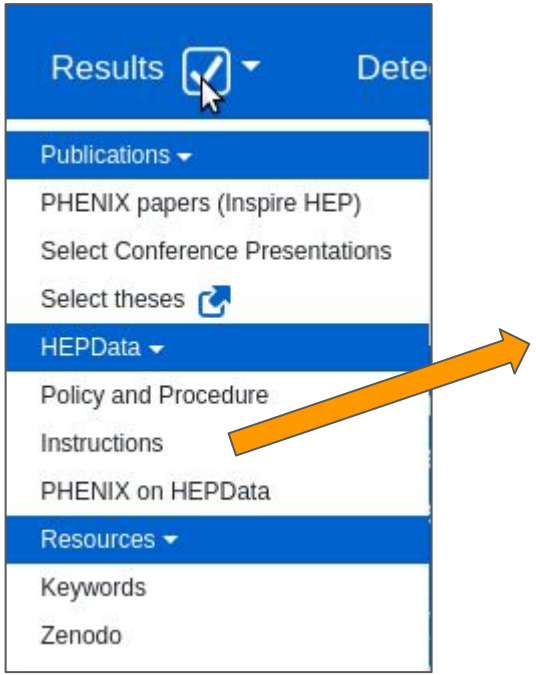
Physics (88 items)

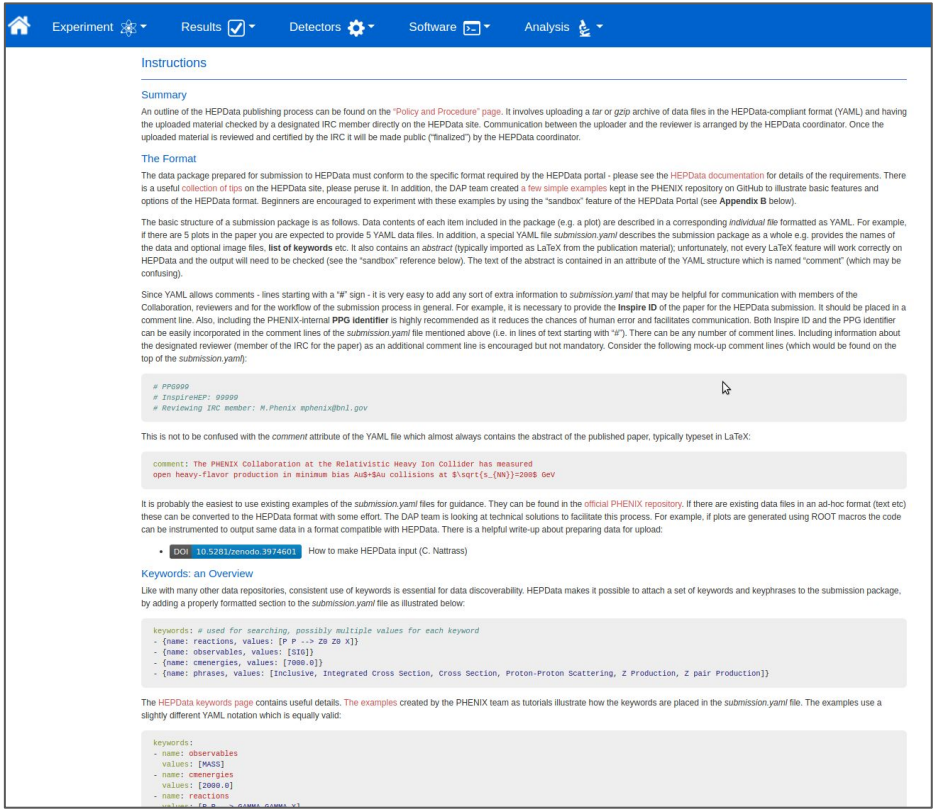
Keyword	Description
3he+au	Helium3-on-gold collisions
anisotropy	Anisotropy
asymmetry	Asymmetry
au+au	Gold-on-gold collisions
azimuthal	Azimuthal
b-meson	B meson
backward-rapidity	The backward kinematic region
binary scaling	Binary scaling
bose-einstein	Bose-Einstein statistics
bottom	Particles containing the b-quark
centrality	Centrality characteristic of the collision
cgc	Color Glass Condensate (type of matter)
charm	Particles containing the c-quark
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
d+au	deutron-on-gold collisions
d-meson	D meson
dca	Distance of Closest Approach

HEPData

- 46 items uploaded i.e. data from 46 published papers (such as used in plots) is now available for reliable download in a standard format, and is indexed and discoverable.
- We are still catching up with other major experiments who started investing in this earlier.
- Once again, DOIs.
- The official publication policy in PHENIX mandates the HEPData process for each potential publication. Steady state work now.
- HEPData/Rivet workshops in Fall 2020 (thanks to Christine for organizing this).
- Documentation: based on our experience the policies and procedures have been discussed and updated.
- Extensive re-write of the corresponding documentation on the website (next page)

HEPData Instructions and Policies

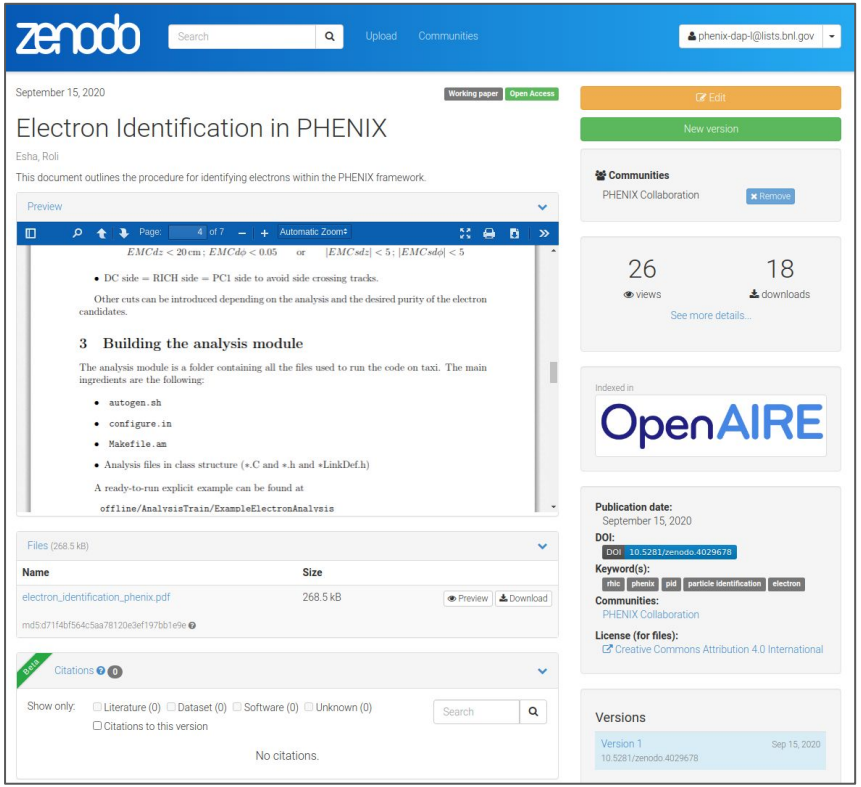




Analysis resources on the site

- Progress has been made with creating updated tutorials for specific work areas (see on the site)
- Currently implemented as write-ups published on Zenodo and linked on the site
- Caveat: these materials need another round of updates to account for mode code access (i.e. update or eliminate references to the code browser only available internally, review placement of code on AFS and whether it's suitable for publication etc)

Example of a tutorial uploaded to Zenodo



Analysis Content on the site

- Having well documented use cases (tutorials and/or complete preserved analyses) is obviously conducive to both efficiencies in medium term and longer term DAP
- Better engagement by the PHENIX community still remains on the critical path
  - Reverse engineering of prior analyses turns out to be close to impossible
  - Analyzers active in a given area are the best resource we have to create materials to meet education and analysis preservation goals
- We have good technology when it comes to DAP implementation (Zenodo, REANA etc) but the actual materials still need to be produced

PHENIX School in 2021

- The website can be leveraged as the portal to the necessary writeups and materials
- How much material can be made reusable?
- Timing/schedule?

REANA

- A framework already used by LHC experiments
- Captures
  - Details of the analysis workflow by utilizing a structured description (graph)
  - Software and environment by means of containerization
- A test instance recently became available at BNL thanks to SDCC, initial testing for PHENIX in progress
- At a minimum, requires Docker expertise
- Still need to evaluate cost/benefit
- Looks promising

OpenData

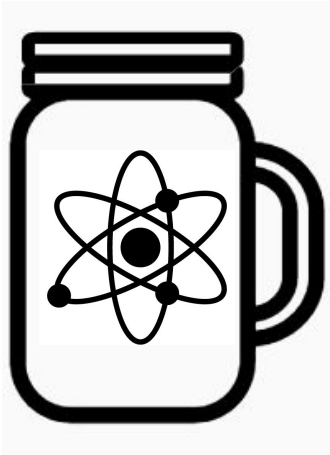
- A CERN portal for capture and publishing of data samples, software and documentation
- Initially reserved for LHC experiments
- PHENIX was given an approval and access in Fall 2020
- Storage allocated
- Currently no cost to PHENIX
- The first package/analysis is currently in the works (thanks Gabor)

Status and Plans

- The new durable DAP website has been commissioned
- Content is being developed and added
- Ongoing creation and uploads of curated and tagged materials to Zenodo
- Systematized submissions to HEPData
- Planned adoption of OpenData and REANA technologies
- Main concern is lack of documented analysis use cases and vetted tutorials
- Preparation for the PHENIX School and DAP are closely related tasks
- Can the EC appoint/name people who would contribute to this effort?

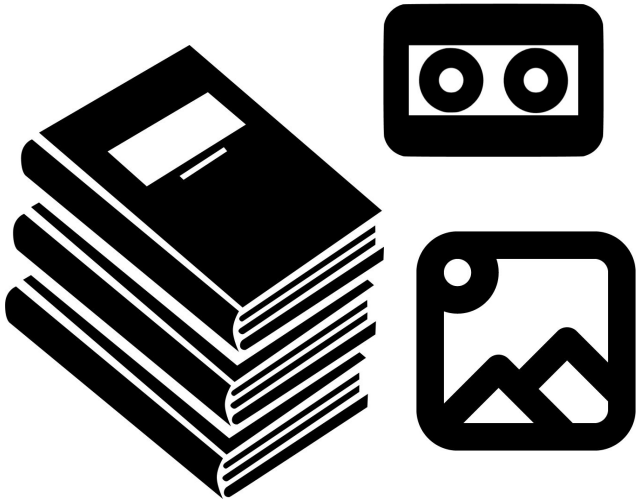
Backup slides

A practical approach to the components of the Knowledge Management in PHENIX



Analysis capture

Web-based documentation



A highly-functional repository for research materials

The website technology

- Inspired by the websites of the HEP Software Foundation and NPPS
  - https://hepsoftwarefoundation.org/
  - https://npps.bnl.gov/
- The website is static (no DB) - using the Jekyll static website generator
  - HTML is generated, not written by hand
  - Easy to contribute (Markdown, YAML) and easy to maintain
  - Data content kept separately from the layout
  - Manipulation of structured data in Jekyll makes for a compact and efficient design of pages, and automated content generation - DB-like functionality at compile time
  - GitHub Pages for management and development + version control + Jekyll builds
  - Portability (easy to export/migrate the complete site as HTML)
  - Performance
  - Security
- The “Bootstrap” toolkit for the layouts/navigation (no custom JS/CSS needed)

Structured data used to generate the site

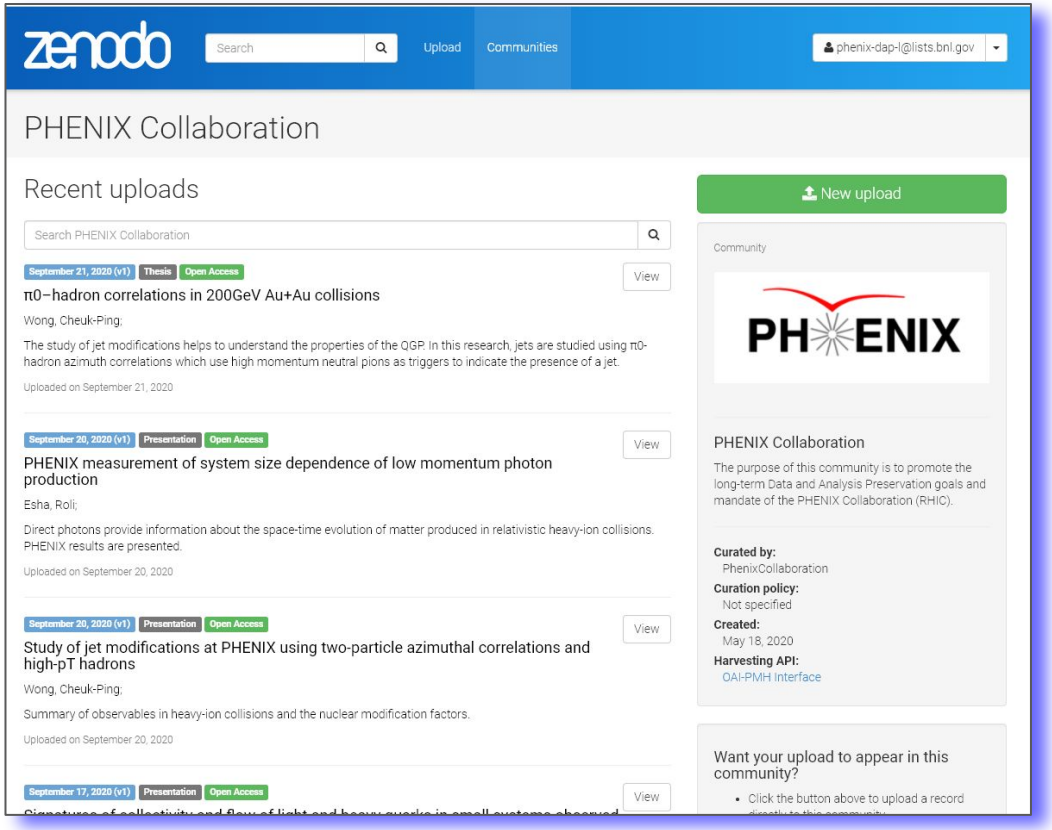
```
############################ RUN 12
- run: run12
  title: Run 12
  period: 2011-2012
  coordinator: Xiaochun He, GSU.
  rhic:
    - {
       species: 'polarized p+p',
       energy: 100.2,
       lumi: '~ /10 <i>pb</i><sup>-1</sup>',
       Nevents:
      }
    - {
       species: 'polarized p+p',
       energy: 254.9,
       lumi: '32/~ <i>pb</i><sup>-1</sup>',
       Nevents:
      }
    - {
       species: '<sup>238</sup>U<sup>92+</sup>+<sup>238</sup>U<sup>92+</sup>',
       energy: 96.4,
       lumi: '0.2<i>nb</i><sup>-1</sup>',
       Nevents: 1.2B/0.8B
      }
    - {
       species: '<sup>63</sup>Cu<sup>29+</sup>+<sup>197</sup>Au<sup>79+</sup>',
       energy: 99.9+100.0,
       lumi: '5<i>nb</i><sup>-1</sup>',
       Nevents: 0.8B/8.1B
      }
    - {
       species: '<sup>197</sup>Au<sup>79+</sup>+<sup>197</sup>Au<sup>79+</sup>',
       energy: 2.5,
       lumi: '-',
       Nevents: Very short
      }
  ert_comment: Summary of thresholds (DAC values). Values in parentheses are for the PbGl.
  ert_thresholds:
    - '02/09/12, 358208, 30(29), 31(30), 29(29), 29(25), 920, Run12pp200 - Pedestal tuned - EMCal dynamic range ~25GeV'
    - '03/20/12, 364957, 30(29), 31(30), 29(29), 29(25), 920, Run12pp510 - EMCal dynamic range ~50GeV'
    - '04/23/12, 369200, 30(29), 31(30), 29(29), 29(25), 920, Run12UU193 - EMCal dynamic range ~25GeV'
    - '05/16/12, 372155, 31(30), 32(31), 30(29), 49(45), 920, Run12CuAu200 - EMCal dynamic range ~25GeV'
```

Leveraging CERN-based tools

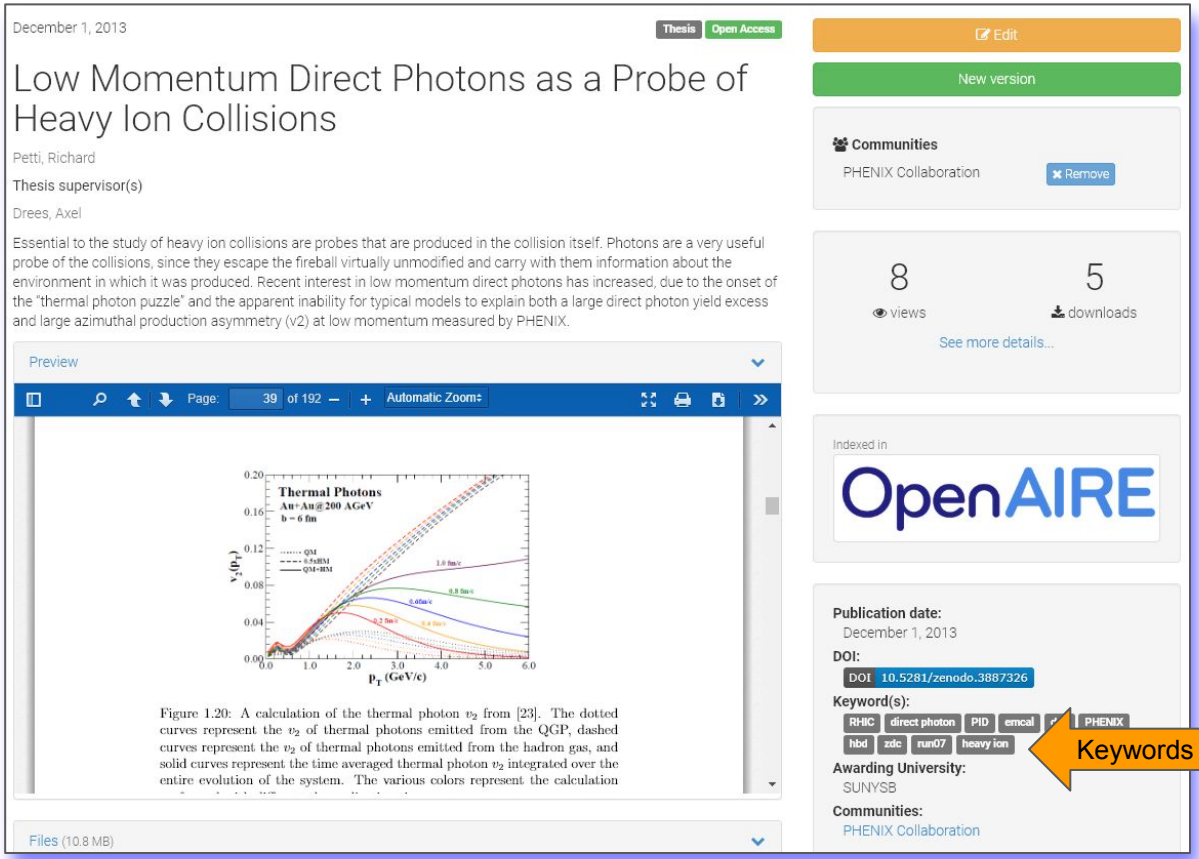
- Using **InspireHEP** to keep the publication list and cross-reference HEPData
- **Zenodo**
  - Based on the Invenio framework created at CERN
  - Active migration of relevant documents from legacy services to Zenodo is under way
  - Using a curated list of keywords for - easy to find physics topics, detector elements, conference materials etc
  - Easy to cite using the DOIs (persistent identifiers)
  - BNL is a collaborator in the InvenioRDM project - the next generation repository
- **HEPData**
  - Based on the Invenio framework
  - Recently renewed effort to create submission packages and commit new data to the service
  - **As of this week, new policy by PHENIX IB in place: HEPData package as a prerequisite for publication**
- The CERN Open Data Portal - material in the works, testing under way
- REANA - initial testing

Zenodo@CERN - the PHENIX community

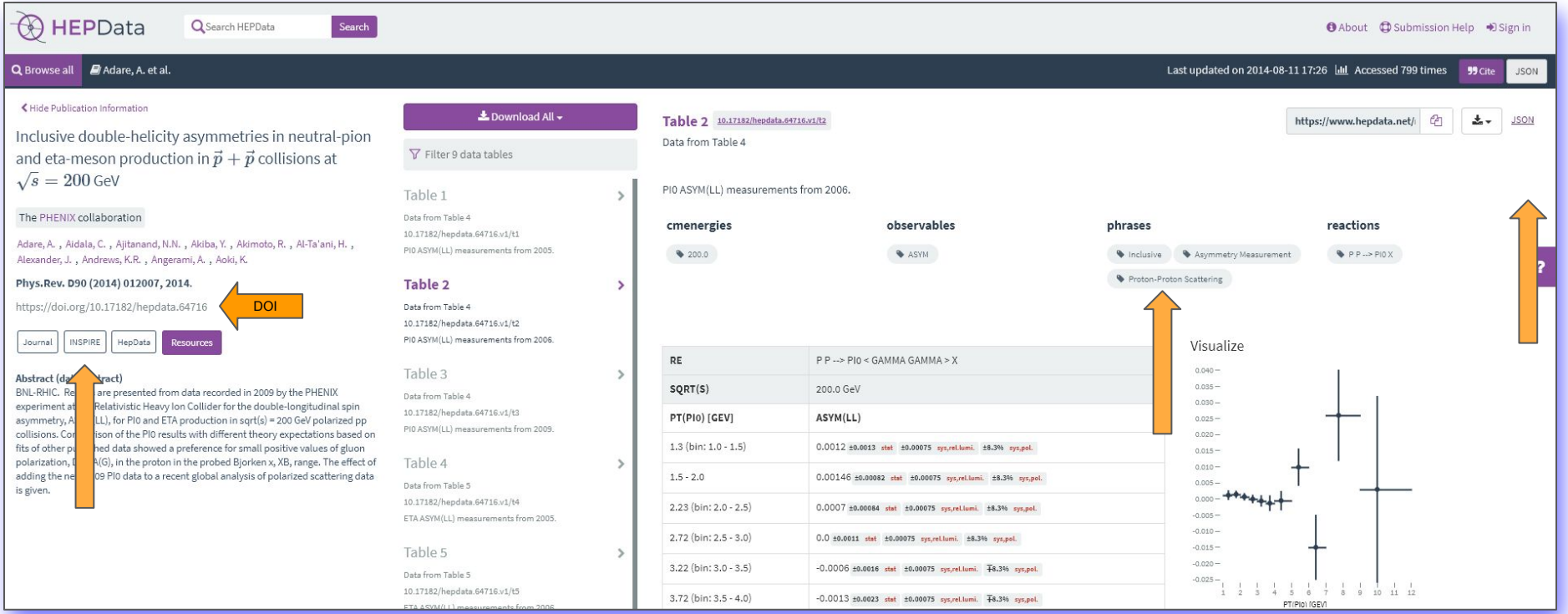
- Branded
- Curated
- Discoverable
- Indexed (keywords)
- Elastic search capability



An example of a PHENIX item on Zenodo

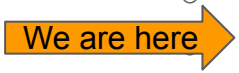


HEPData: an example of a PHENIX entry



Open Data Definitions

- “Canonical” data levels
  - Level 1 data provides more information on published results
  - Level 2 data includes simplified data formats for outreach and analysis training
  - Level 3 data comprises reconstructed collision data and simulated data together with analysis-level experiment-specific software
  - Level 4 data covers basic raw data
- Where does PHENIX stand?
  - Level 1 is covered by the current HEPData activity and auxiliary info committed to Zenodo
  - Level 2: work underway to create Ntuples illustrating analysis techniques *(next slide)*
  - Level 2: the Open Data Portal *(under discussion)*
  - Levels 3 and 4 are not practical for public access for much of the same reasons as exist in other experiments (e.g. access to calibrations and access to sites)



The Open Data effort: annotated Ntuples (G. David)

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 oft... 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 cluste...
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

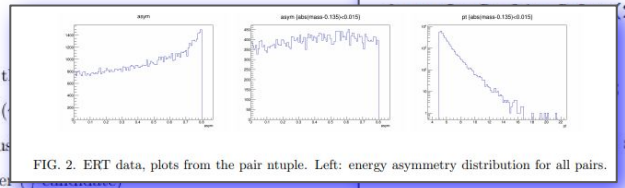


FIG. 2. ERT data, plots from the pair ntuple. Left: energy asymmetry distribution for all pairs.

```
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);
```

see Fig. 1.

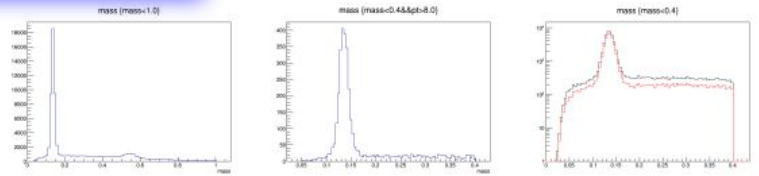


FIG. 1. ERT data, plots from the pair ntuple. Left: Invariant mass in the 0-1 GeV region. You can see a strong π^0 and a well-recognizeable η peak. Middle: π^0 peak for pairs with p_T greater than 8 GeV/c. You can clearly see the combinatorial background outside the peak, which should

Ntuples: O(100MB) each
Hosting options: Zenodo, Open Data