Data & Analysis Preservation and Open Data: Experience in PHENIX

Maxim Potekhin
Nuclear and Particle Physics Software Group

BROOKHAVEN NATIONAL LABORATORY

Future Trends in Nuclear Physics Computing

09/30/2020

PHENIX

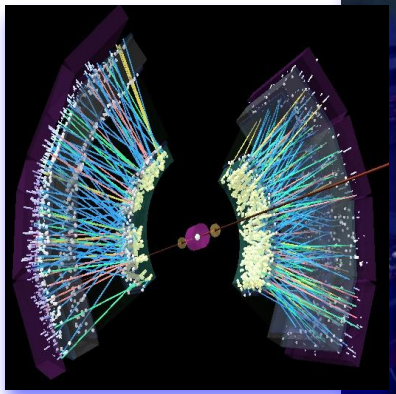
This presentation

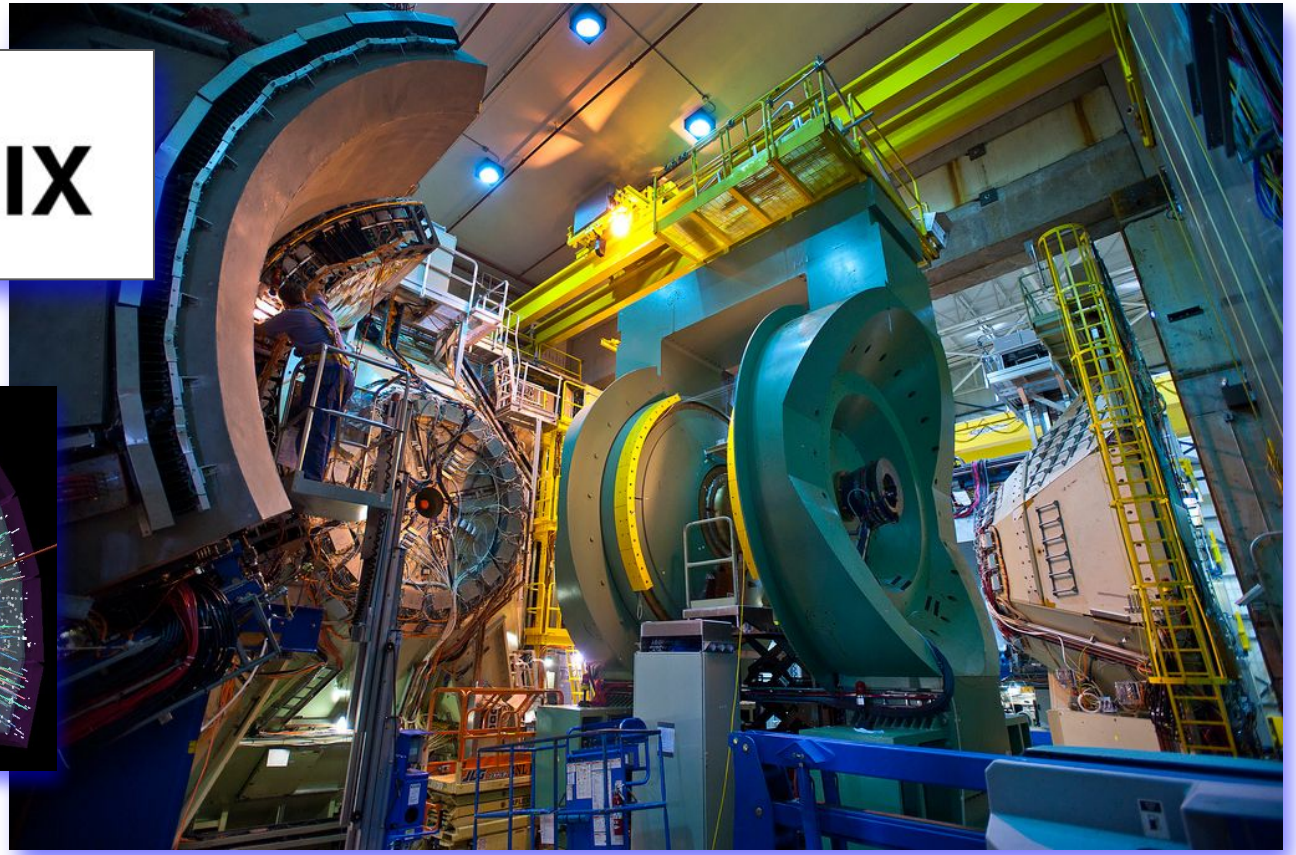
- The PHENIX Collaboration became actively engaged in Data and Analysis Preservation (DAP) in 2019 and is facing challenges in that area common to other experiments (e.g. at RHIC, JLab and elsewhere)
- The goal of today's presentation is to consider ways of leveraging the PHENIX DAP experience for the benefit of the community and identifying venues for collaboration in that area

An Overview

- PHENIX in a nutshell
- Data and Analysis Preservation (DAP)
- DAP: various aspects and challenges in PHENIX
- Technical solutions identified, leveraged or developed for the PHENIX DAP
- Lessons learned + potential for collaboration

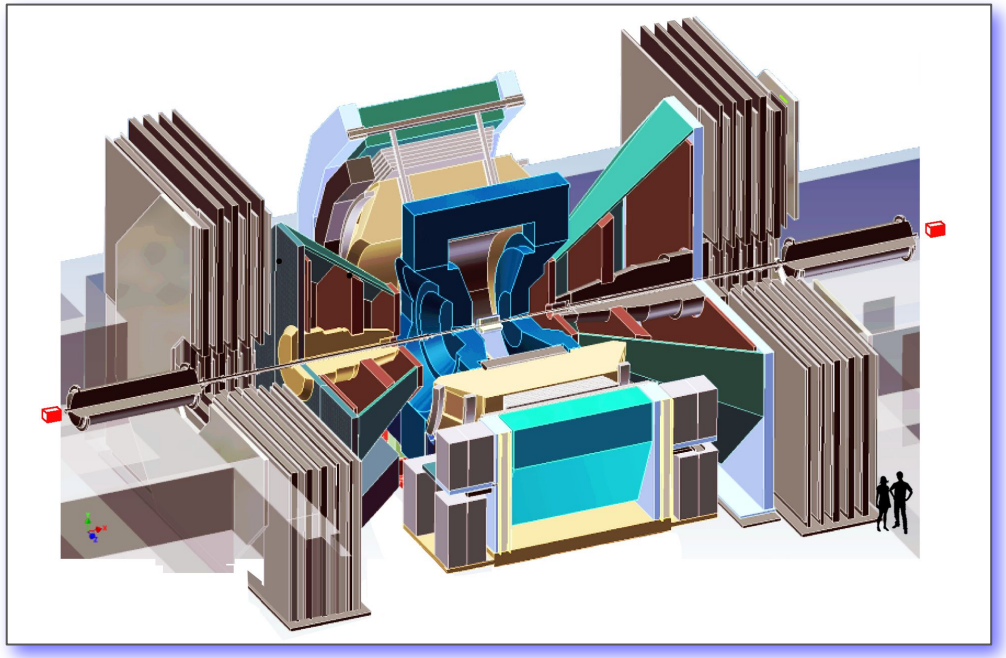
PHENIX





PHENIX in a nutshell

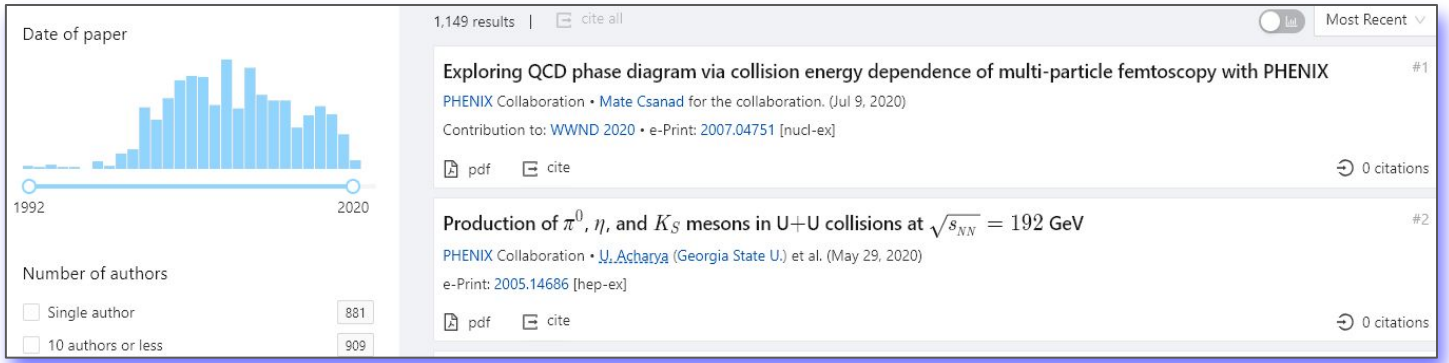
- “Pioneering High Energy Nuclear Interaction eXperiment”
- One of the two large RHIC experiments
- A large, complex general purpose detector with a considerable physics reach
  - Search for QGP and study of its properties, spin physics, other topics
  - Complex analyses



- 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
  - A total of 165 PhD theses and counting
- Key active analyses considered a priority (effort limited, need to prioritize):
  - Heavy flavor in central and forward rapidity ranges
  - Low pT photons and thermal dileptons
  - High pT direct photon and photon-jets



Definition of the problem as per the “Software Preservation and Legacy issues at LEP” (J.Shiers)

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

DAP

- DAP = Data and **Analysis** Preservation
  - A tremendous investment which needs to be preserved
  - Potential to provide new insights, validate new models, possibly perform cross-checks etc
  - Synergy with near and medium term goal of maintaining quality of analyses
- Recognized as an important work area by the HEP and NP communities
  - Solid coverage at CHEP 2019 and recent workshops (CERN and others)
  - cf. HSF Community White Paper https://arxiv.org/abs/1810.01191 and other papers
  - DAP has become a standard item in the experiment and facility reviews
  - However, funding can be a problem - DAP requires additional organizational, software, documentation and other effort
  - While “cost justification” is generally acknowledged such justification does not always result in action and/or funding
- DAP implementation = *Infrastructure and facility + Knowledge Management*

Data preservation: the role of the facility

- DAP universally depends on continuity of services and expertise provided by the facility - this is especially true in PHENIX case where BNL SDCC is the only fully-capable computing site for that experiment.
- In addition to bit preservation (mass storage) the facility provides software building and provisioning capabilities (including containers, CVMFS etc), databases and much more.
- In terms of planning, the facility involvement should cover the time period relevant to a particular DAP effort. Cf. if there is a dependence on AFS it must be somehow satisfied over the years.

Facility contribution to the community infrastructure

- BNL SDCC is a collaborator in the InvenioRDM project
  - BNL lead: C. Gamboa
- A local instance of Zenodo at BNL has been deployed
- ...more details on Zenodo below

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.

PHENIX analysis Knowledge Management: software and data

- Software provenance, dependencies, APIs and configuration (e.g. which specific macros were run, with what arguments, in what sequence and under what conditions, linked to what external libraries etc)
- “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 only to the people involved in this analysis
  - 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)
- 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

A legacy Knowledge Management tool: analysis notes

- By policy, each analysis progressing towards eventual publication is expected to be documented in an “analysis note” (similar to other experiments)
  - This is a core instrument to ensure the quality of analyses
- In addition to the physics content each note *should* contain enough information for a qualified physicist to reproduce the results
  - Description and location of (archived) macros and other software
  - Description and location of the data
  - Other data and software dependencies (additional configuration and conditions-type data etc)
- This is codified in the “analysis note template” and policies related to its use
- Analysis code is saved as a tarball (and in CVS for some components)
  - But not really designed for reusability
  - Most of the time no detailed instructions for software configuration and use are provided
- In summary, the analysis notes are not the ideal tool - unless there is an enforceable policy and they become part of the experiment’s culture early on

Knowledge Management: core software documentation

- “PISA” - the core simulation framework in PHENIX (based on Geant3)
- Fun4All - the principal reco framework
- In general, the existing PHENIX documentation for both components is obsolete and existing examples/tutorials are unreliable
- Information is spread over a few Web resources
- This is a work area in itself

Knowledge Management: the detector and its subsystems

- Detailed and evolving experiment configuration information is necessary to maintain the analysis capability
  - Hardware configuration/installation geometry, components upgraded, added or removed
  - Trigger configuration (e.g. EMCAL trigger masks), luminosity, various efficiencies, statistics etc
  - Hard-to-find information e.g. limits on mechanical precision of the detector alignment
- Much of this info spread over multiple Web resources e.g. run pages, Wikis, CAD pages, notes etc - information becomes diluted
- PhD theses are an important component of the documentation, and until recently they were stored in a proprietary DB on the PHENIX server
  - Now being moved to Zenodo (CERN) - complex queries are now possible
- Other detector-related references (NB. need to be mindful of copyright issues)

The HSF Community White Paper: “Data and Software Preservation to Enable Reuse”

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

An example of a working directory used in an analysis.

Reverse engineering is difficult.

```
check_pdst_vertexfile.root
condor_bb.job
condor_cc.job
condor_hadrons.job
condor.job
condor_jpsi.job
condor_ktomu.job
condor_step3.job
corruptfiles
CorruptFiles_bb.txt
CorruptFiles_cc.txt
CorruptFiles_hadrons.txt
CorruptFiles_jpsi.txt
corruptfiles.txt
Corruptfiles.txt
dcar_bbblack_jpsired_pdst.pdf
dcar_bbblack_jpsired_pdst_ylog.pdf
dcar_bb_from_pythia_logy.pdf
dcar_bb_from_pythia.pdf
dcar_jpsi_from_pythia_logy.pdf
dcar_jpsi_from_pythia.pdf
DeadChannels.dat
diff_mydir_cesardir.txt
DST_embed_bb
DST_embed_bbbar
DST_embed_cc
DST_embed_hadrons
DST_embed_hadrons_3gev
DST_embed_jpsi
DST_embed_ktomu
DST_pythia_hadrons
dst_pythia_hadrons-MB0-0000429896-0000.list
dstpythia.list
dstpythia.root
dstpythia.txt
DST_sim_bb
DST_sim_bbbar
DST_sim_cc
DST_sim_hadrons
DST_sim_hadrons_3gev
DST_sim_hadrons_combined
DST_sim_hadrons_mineXuanSample
DST_sim_jpsi
DST_sim_ktomu
embedpythia.pdf
env.log
env.txt
ERR
event_gen_bb.csh
event_gen_bb_step2.csh
event_gen_bb_step3.csh
event_gen_cc.csh
event_gen.csh
event_gen.csh~
event_gen_hadrons.csh
event_gen_hadrons_step1.csh
event_gen_hadrons_step2.csh
event_gen_hadrons_step3_reass.csh
event_gen_hadrons_step3_xuan.csh
event_gen_jpsi_combined.csh
event_gen_jpsi.csh
event_gen_ktomu.csh
geom.root
geom_run15_v2.root
get_nevents
get_nevents.C
hadd_files.sh
HV
jpsipdst.txt
landau_parameters.txt
libpicodst_object.so
listofcc.txt
listofpythia.txt
LOG
logall.txt
logfiles.txt
logfile.txt
loglog.txt
logout.txt
logs.txt
log.txt
make_picodstobj.C
mc_bjpsi.C
mc_bjpsi_C_ACLiC_dict_rdict.pcm
mc_bjpsi_C.d
mc_bjpsi_C.so
mc_bjpsi.h
merge_picodsts.C
model_bdecay.C
muid_tube_eff_north_Run15pp200.txt
muid_tube_eff_south_Run15pp200.txt
mut.disabledWires.dat
mvcorrputfiles_bb.sh
mvcorruptfiles.sh
mylogfiles.txt
mypythia_cmu_pdst.pdf
mypythia.pdf
mypythia_pdst.pdf
mysimhad.txt
nokeys.txt
Nonexistingfiles.list
Nonexisting.txt
nooffiles.txt
notopen_bb_1.txt
notopen_bb_2.txt
notopen_bb_3.txt
notopen_bb_4.txt
notopen_bb_5.txt
notopen_bb.txt
NotOpenfiles.txt
notopen_hadrons_ajeeta.list
notopen_jpsi_1.txt
notopen_jpsi_2.txt
notopen_jpsi_3.txt
notopen_jpsi_4.txt
notopen_jpsi.txt
notopen.txt
Old
old_pDST_bb
out.log
output_bb_step2.txt
output_bb_step3.txt
output_bb.txt
output_env.txt
PISA
PisaFullfiles.txt
pisa.txt
practice.list
pylist12.dat
#pythia_bb.C#
pythia_bb.C
pythia_bb.C~
pythia_bb_C_ACLiC_dict_rdict.pcm
pythia_bb_C.d
pythia_bb_C.so
pythia_bb.h
pythia_bb.h~
pythia_configuration
pythia_files_bb
pythia_files_bbbar
pythia_files_bb_iter0
pythia_files_bb_list.txt
pythia_files_cc
pythia_files_hadrons
pythia_files_hadrons_3gev
pythia_files_hadrons_combined
pythia_files_hadrons_list.txt
pythia_files_hadrons_mineXuanSample
pythia_files_jpsi
pythia_files_jpsi_combined
pythia_files_jpsi_list.txt
pythia_files_ktomu
pythiafiles.txt
pythia.txt
realdataBG-run15pp_file-forAjeeta.list
realdataBG-run15pp_file-forAjeeta_onefile.list
realdataBG-run15pp_file-forCesar.list
realdataBG-run15pp_file-forXuan.list
realdataBG-run15pp_file.list
realdataBG-run15pp_file_split-forAjeeta1.list
realdataBG-run15pp_file_split-forAjeeta2.list
realdataBG-run15pp_file_split-forAjeeta.list
realdataBG-run15pp_file_split-forCesar1.list
realdataBG-run15pp_file_split-forCesar2.list
realdataBG-run15pp_file_split-forCesar.list
realdataBG-run15pp_file_split-forXuan1.list
realdataBG-run15pp_file_split-forXuan2.list
realdataBG-run15pp_file_split-forXuan.list
realdataBG-run15pp_file_split.list
Run_reassociation.C
run_segments.list
Sim3D++.root
SimFileLists
singlemuon_embed_pdst.root
singlemuon_jpsi_embed_pdst-MB0-0000422070-0001-2.root
split.csh
split_dsts.C
splitfiles
svxPISA.par
temp.txt
totest.list
whencreated.txt
with_n_option.txt
Xuan_allfailedFiles.txt
xuanfilelist.txt
```

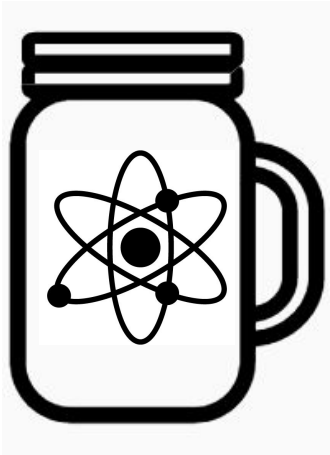
Challenges in PHENIX

- “Retrofitting is nearly impossible” (as it concerns various DAP tools)
  - So the next best thing is an accurate record of the analysis procedure and its components
- Diminishing manpower
  - Progressively smaller teams at each participating institution
- Dissipation of expertise, combined with insufficient maintenance of many types of documentation and its fragmentation
- Analyses still ongoing - pressure on PhD students, postdocs and other personnel to get it done with modest available resources
  - DAP requirements being a significant additional hurdle, hard to find reliable volunteers
  - Currently, making an extra effort is not rewarded in any meaningful way (see next slide)
- New students need training and solid information sources to learn from
- The PHENIX software is not easily portable

DAP and Software Sustainability

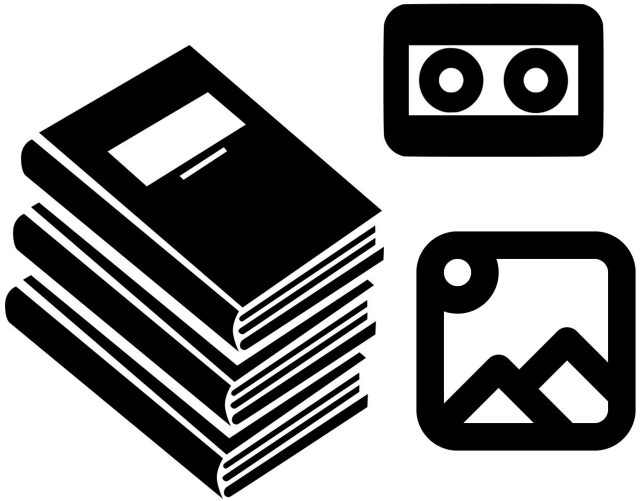
- A few points from the presentation by Daniel Katz at this workshop on 9/29/2020 are quite relevant for DAP, with regards to the analysis software
- Quote:
  - *Software development and maintenance is human-intensive*
  - *Much software developed specifically for research, by researchers*
  - *Researchers know their disciplines, but often not software best practices*
  - *Researchers are not rewarded for software development and maintenance in academia*
- The word “incentive” is used 3 times in Daniel’s talk
- Incentivizing could be the key to changing the culture - DAP is serious work and it needs recognition

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 Analysis Environment

- Core analyses take place within the SDCC facility at BNL
- AFS is used for software provisioning (regular builds etc)
- SL7
- A large part of the analysis code is managed in CVS (legacy)
- Some of the conditions data is kept as files accessed via scripts
  - Part of the reason for creation of custom data products managed in individual analyses
- The cost/benefit situation in PHENIX currently does not favor creation of containers for analysis purposes
  - Containers are used in production for separate reasons

Analysis Preservation

- In practice we can't capture all - or many - analyses for the following reasons
  - PHENIX started >20 years ago with no defined DAP policy, plan or framework
    - **this changed in 2019**
  - Once a postdoc or a graduate student leaves, there is little chance of knowledge recovery
  - Only current/recent analyses are viable for preservation
- There is an overlap between medium term analysis capability and longer term analysis preservation - an extra motivation for DAP right now
- *Creating up-to-date summaries of analysis techniques (e.g. PID etc) is universally helpful - progress is being made*

Web-based documentation: the new PHENIX website

- https://www.phenix.bnl.gov/
- The hub for all DAP information, designed to be a long-term resource
- Completely new layout and interface
- Content consolidated from a number of legacy web resources
- Eliminates in-house DB management, minimizes local storage and instead relies on external services for hosting content:
  - Zenodo
  - HEPData
  - GitHub
- Where possible, relies on DOIs for references, to achieve long-term integrity

The website technology

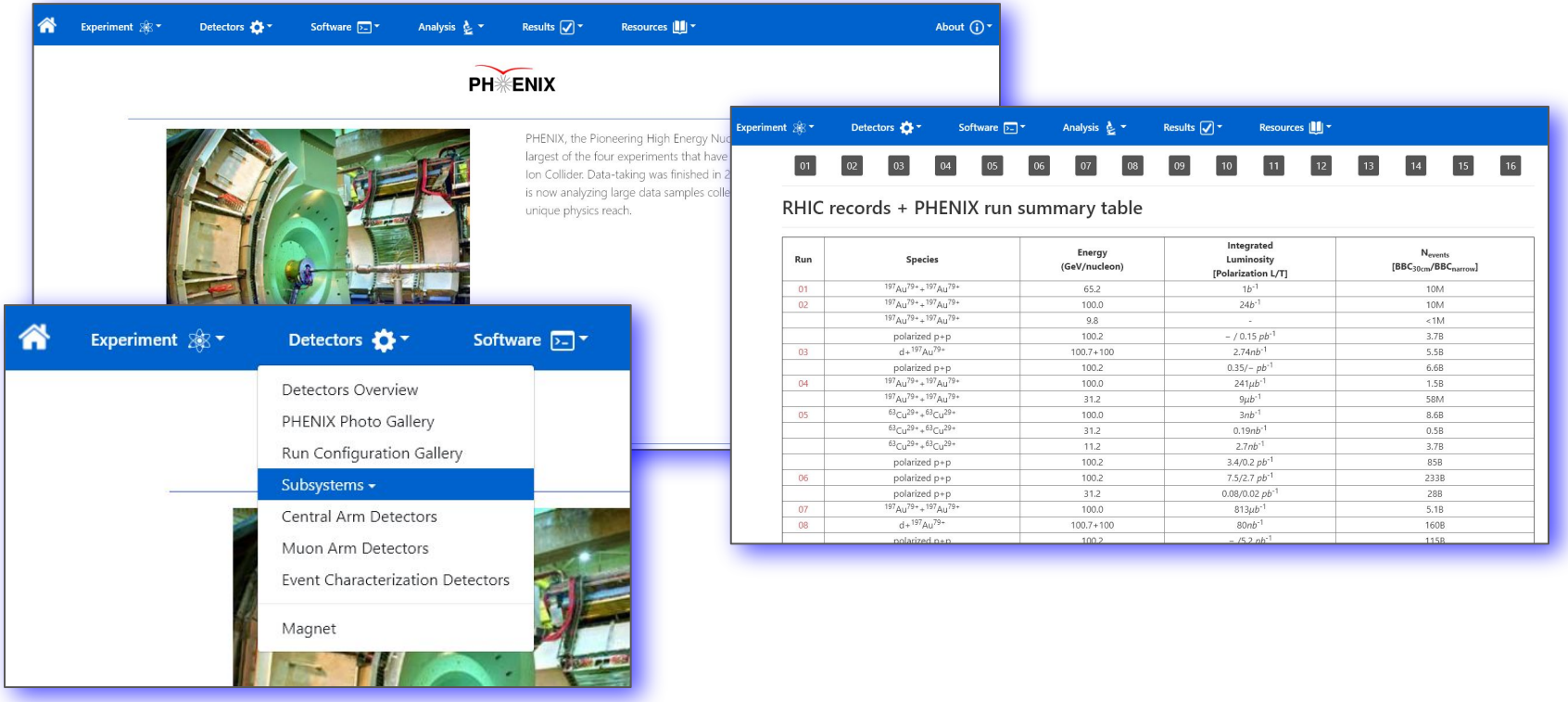
- Inspired by the websites of the HEP Software Foundation and NPPS
  - https://hepsoftwarefoundation.org/
  - https://npps.bnl.gov/
- The website is static (no DB) - using the Jekyll static website generator
  - HTML is generated, not written by hand
  - Easy to contribute (Markdown, YAML) and easy to 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'
```

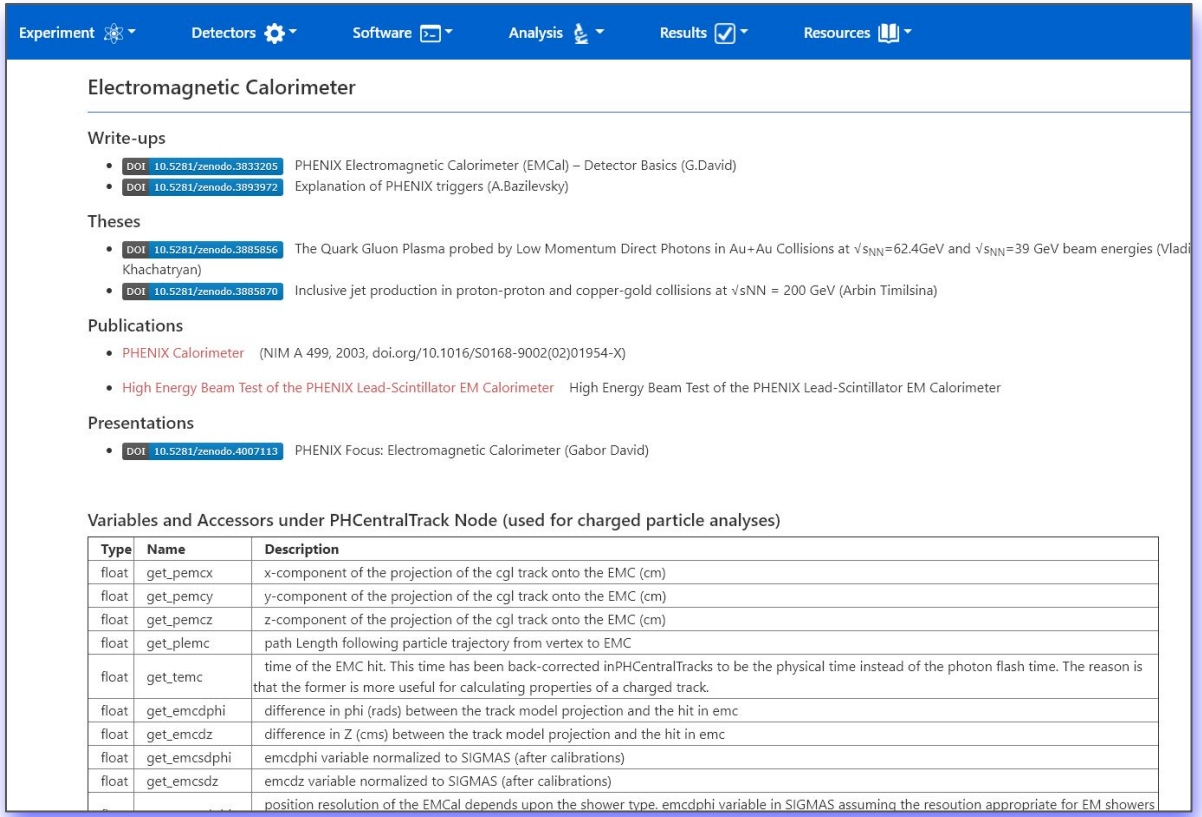
The new PHENIX website

https://www.phenix.bnl.gov/

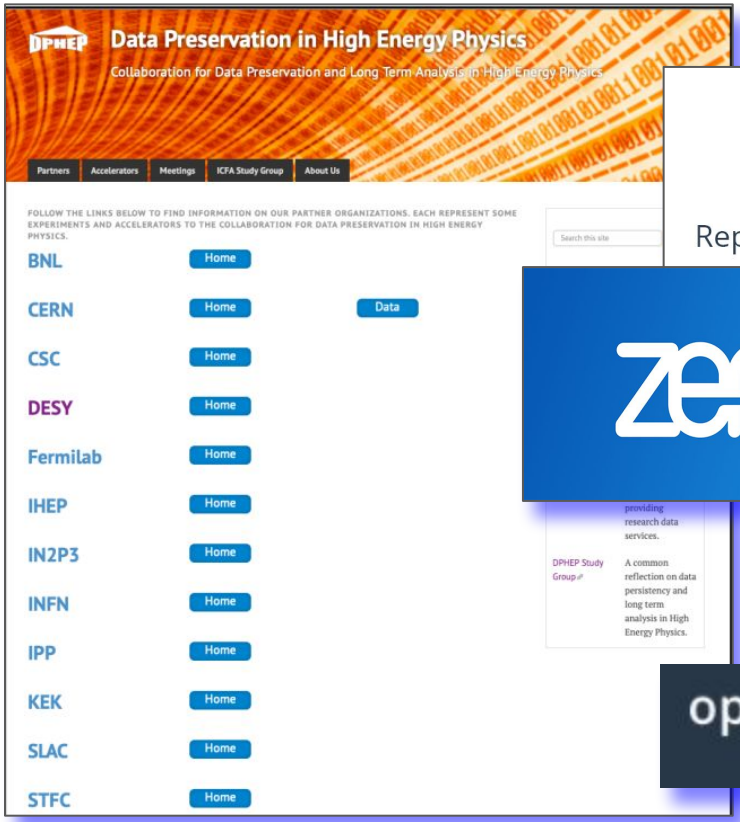


Run	Species	Energy (GeV/nucleon)	Integrated Luminosity [Polarization L/T]	N_{events} [BBC_{30cm}/BBC_{narrow}]
01	$^{197}Au^{79+}$+$^{197}Au^{79+}$	65.2	$1b^{-1}$	10M
02	$^{197}Au^{79+}$+$^{197}Au^{79+}$	100.0	$24b^{-1}$	10M
	$^{197}Au^{79+}$+$^{197}Au^{79+}$	9.8	-	<1M
	polarized p+p	100.2	– / 0.15 pb^{-1}	3.7B
03	d+$^{197}Au^{79+}$	100.7+100	$2.74nb^{-1}$	5.5B
	polarized p+p	100.2	0.35/– pb^{-1}	6.6B
04	$^{197}Au^{79+}$+$^{197}Au^{79+}$	100.0	$241\mu b^{-1}$	1.5B
	$^{197}Au^{79+}$+$^{197}Au^{79+}$	31.2	$9\mu b^{-1}$	58M
05	$^{63}Cu^{29+}$+$^{63}Cu^{29+}$	100.0	$3nb^{-1}$	8.6B
	$^{63}Cu^{29+}$+$^{63}Cu^{29+}$	31.2	$0.19nb^{-1}$	0.5B
	$^{63}Cu^{29+}$+$^{63}Cu^{29+}$	11.2	$2.7nb^{-1}$	3.7B
	polarized p+p	100.2	3.4/0.2 pb^{-1}	85B
06	polarized p+p	100.2	7.5/2.7 pb^{-1}	233B
	polarized p+p	31.2	0.08/0.02 pb^{-1}	28B
07	$^{197}Au^{79+}$+$^{197}Au^{79+}$	100.0	$813\mu b^{-1}$	5.1B
08	d+$^{197}Au^{79+}$	100.7+100	$80nb^{-1}$	160B
	polarized p+p	100.2	– /5.2 pb^{-1}	115B

An example of a subsystem page



CERN and community tools for DAP



reana

Reproducible research data analysis platform

HEPData

zenodo

iNSPIRE HEP

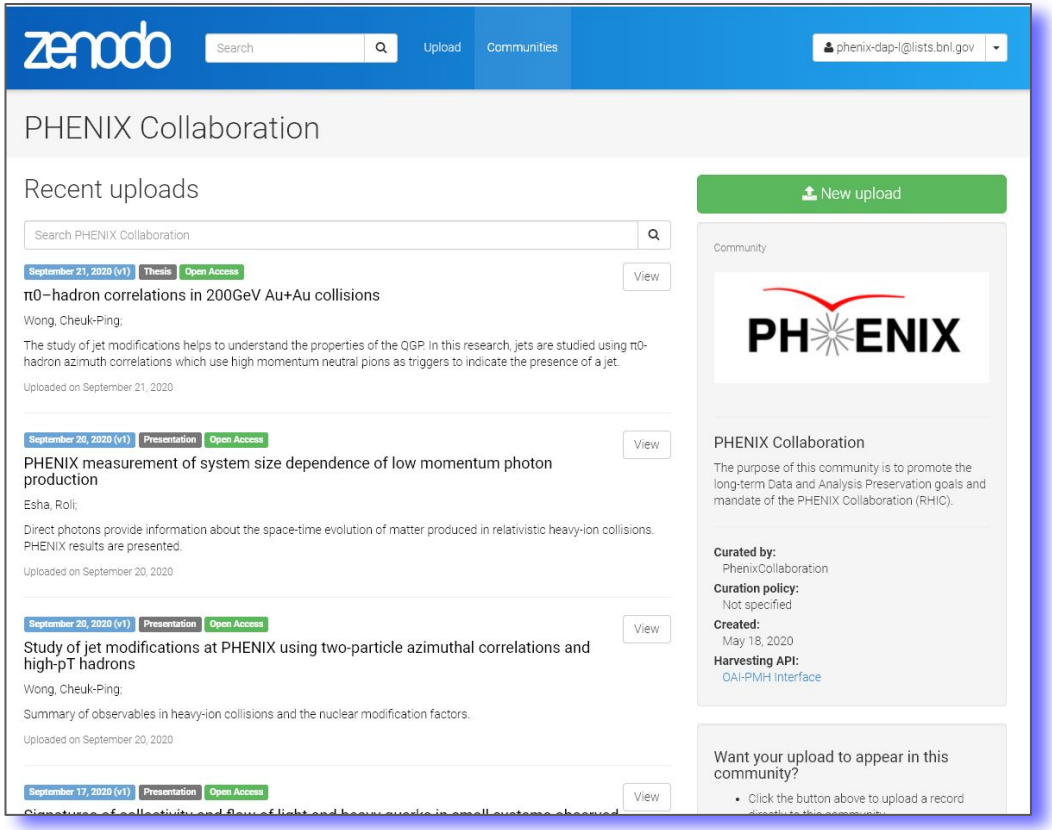
opendata CERN

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 - access under discussion, material in the works
- REANA, Rivet and other tools
  - Use of these tools is only possible with highly organized and documented active analyses
  - Retrofitting is quite hard, limited available effort likely means these tools won't be used
  - Instead, annotate/document analyses (in progress)

Zenodo@CERN - the PHENIX community

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



An example of a PHENIX item on Zenodo

December 1, 2013

Thesis | Open Access

Low Momentum Direct Photons as a Probe of Heavy Ion Collisions

Petti, Richard

Thesis supervisor(s)

Drees, Axel

Essential to the study of heavy ion collisions are probes that are produced in the collision itself. Photons are a very useful probe of the collisions, since they escape the fireball virtually unmodified and carry with them information about the environment in which it was produced. Recent interest in low momentum direct photons has increased, due to the onset of the "thermal photon puzzle" and the apparent inability for typical models to explain both a large direct photon yield excess and large azimuthal production asymmetry (v2) at low momentum measured by PHENIX.

Preview

Page: 39 of 192 — + Automatic Zoom

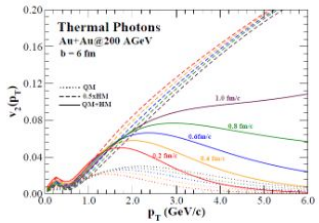


Figure 1.20: A calculation of the thermal photon v_2 from [23]. The dotted curves represent the v_2 of thermal photons emitted from the QGP, dashed curves represent the v_2 of thermal photons emitted from the hadron gas, and solid curves represent the time averaged thermal photon v_2 integrated over the entire evolution of the system. The various colors represent the calculation

Files (10.8 MB)

Edit

New version

Communities

PHENIX Collaboration — Remove

8 views

5 downloads

See more details...

Indexed in OpenAIRE

Publication date:
December 1, 2013

DOI:
DOI 10.5281/zenodo.3887326

Keyword(s):
RHIC | direct photon | PID | emcal | PHENIX | hbd | zdc | run07 | heavy ion

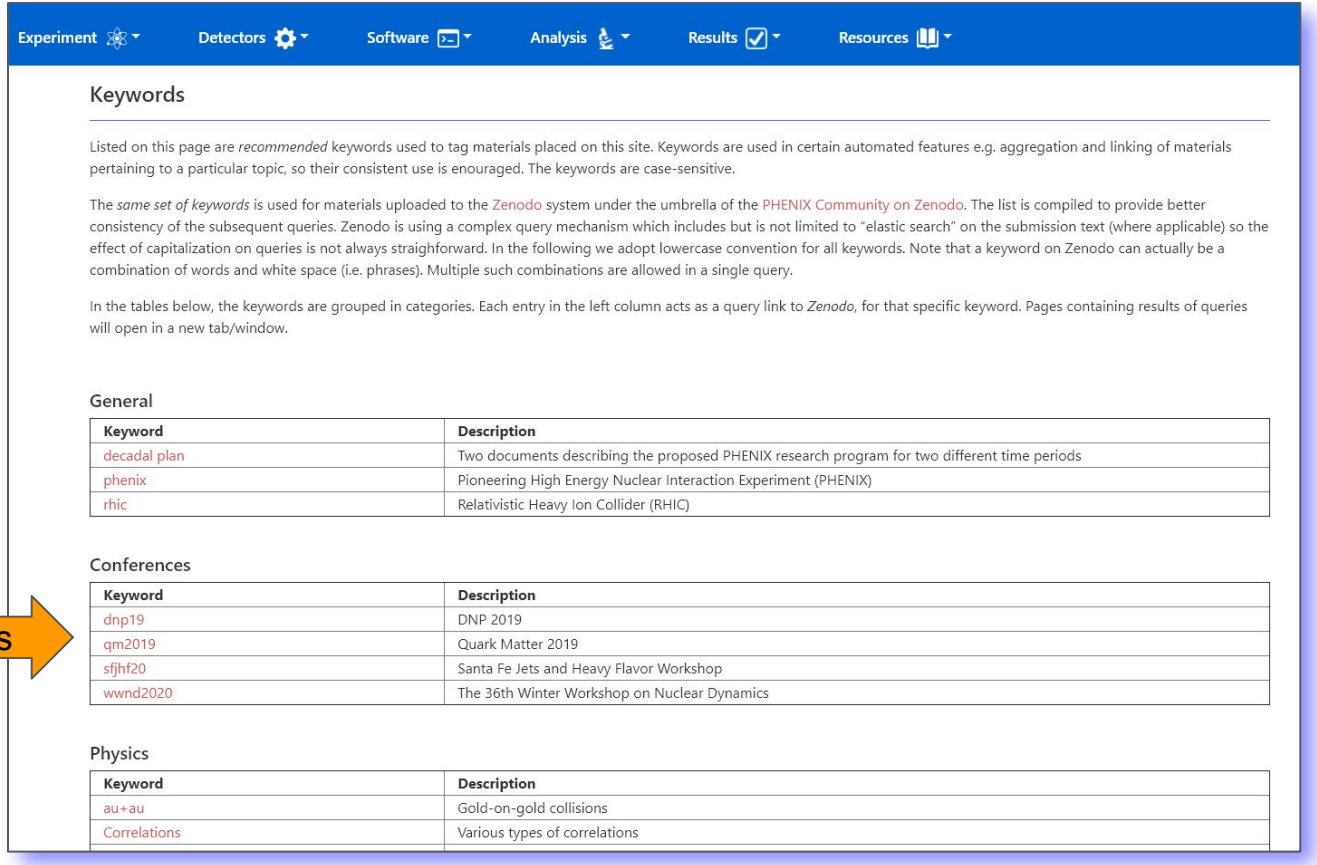
Keywords

Awarding University:
SUNYSB

Communities:
PHENIX Collaboration

Managing the keywords

Working Zenodo links



Experiment | Detectors | Software | Analysis | Results | Resources

Keywords

Listed on this page are *recommended* keywords used to tag materials placed on this site. Keywords are used in certain automated features e.g. aggregation and linking of materials pertaining to a particular topic, so their consistent use is enouraged. The keywords are case-sensitive.

The *same set of keywords* is used for materials uploaded to the Zenodo system under the umbrella of the PHENIX Community on Zenodo. The list is compiled to provide better consistency of the subsequent queries. Zenodo is using a complex query mechanism which includes but is not limited to "elastic search" on the submission text (where applicable) so the effect of capitalization on queries is not always straighforward. In the following we adopt lowercase convention for all keywords. Note that a keyword on Zenodo can actually be a combination of words and white space (i.e. phrases). Multiple such combinations are allowed in a single query.

In the tables below, the keywords are grouped in categories. Each entry in the left column acts as a query link to *Zenodo*, for that specific keyword. Pages containing results of queries will open in a new tab/window.

General

Keyword	Description
decadal plan	Two documents describing the proposed PHENIX research program for two different time periods
phenix	Pioneering High Energy Nuclear Interaction Experiment (PHENIX)
rhic	Relativistic Heavy Ion Collider (RHIC)

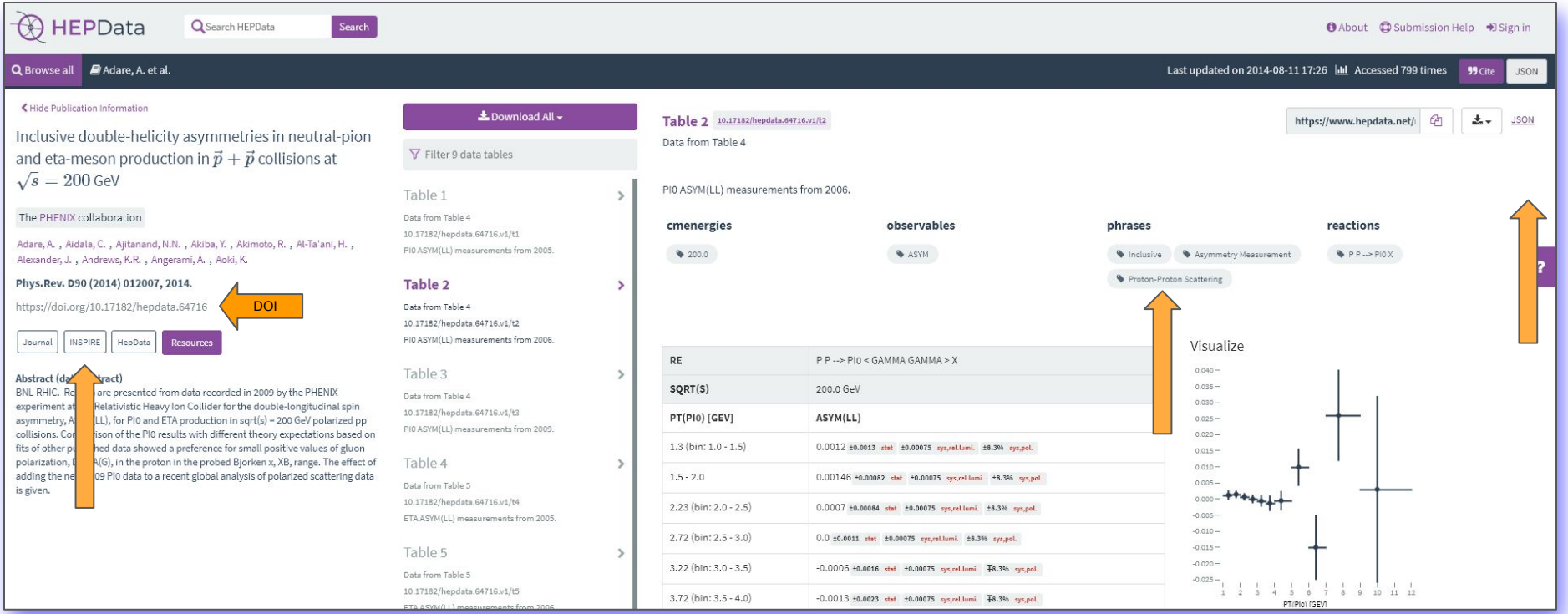
Conferences

Keyword	Description
dnp19	DNP 2019
qm2019	Quark Matter 2019
sfjhf20	Santa Fe Jets and Heavy Flavor Workshop
wwnd2020	The 36th Winter Workshop on Nuclear Dynamics

Physics

Keyword	Description
au+au	Gold-on-gold collisions
Correlations	Various types of correlations

HEPData: an example of a PHENIX entry



Open Data

- “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 z 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 clu
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

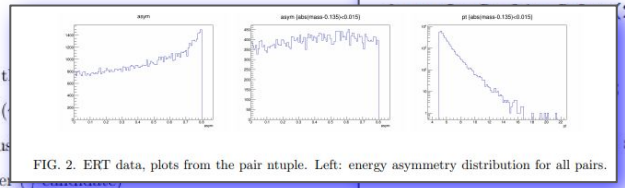


FIG. 2. ERT data, plots from the pair ntuple. Left: energy asymmetry distribution for all pairs.

```
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.

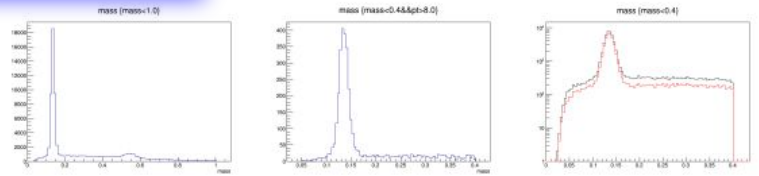


FIG. 1. ERT data, plots from the pair ntuple. Left: Invariant mass in the 0-1 GeV region. You can see a strong π^0 and a well-recognizeable η peak. Middle: π^0 peak for pairs with p_T greater than 8 GeV/c. You can clearly see the combinatorial background outside the peak, which should

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

Status

- Infrastructure for the PHENIX DAP has been put in place
- Analysis capture is limited to annotation and documentation of a few use cases due to insufficient effort available and difficulty of retrofitting new structured solutions
- Within the scope described above, PHENIX relies on state-of-the-art services, platforms and tools
- The new durable DAP website has been commissioned and content is being added
- Ongoing creation and uploads of curated and tagged materials to Zenodo
- Systematized submissions to HEPData

Lessons learned

- DAP: plan and start early (should be a part of someone's job description)
- Avoid building in-house information systems wherever possible
  - State-of-the-art services such as Zenodo, HEPData, Inspire etc which cover a vast majority of the experiment's needs when properly used
- Create websites for the long haul: **the content needs to be curated**
  - Information gets fragmented, redundant and/or obsolete (many examples with Wikis etc)
  - Having misleading information is worse than none at all
- A flexible and easy-to-use Conditions DB is key to reducing the amount of ad-hoc data produced and utilized in analyses, hence facilitates DAP
- Prioritize analyses (even with good effort, we won't be able to preserve all)
- Once DAP-friendly frameworks and practices are established in the project the cost of DAP will likely be incremental due to economies of scale and available expertise within the research teams

Potential for Collaboration and Cooperation

- Community workshops and exchange of experience (Rivet/HEPdata workshops planned by the PHENIX community)
  - https://indico.bnl.gov/event/8840/overview
- PHENIX will be happy to share the DAP website development experience
  - Already taking place with shared components of the EICUG Software Group website
- Design of common analysis workflow templates that would be conducive to the analyses capture (cf. REANA)
- Conditions DB
- In future: developing shared expertise in advanced tools like REANA etc