Data & Analysis Preservation: Recent Experience and Implications for the EIC

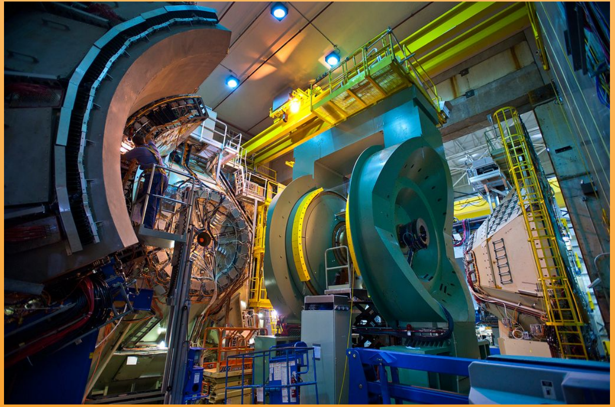
Maxim Potekhin

Nuclear and Particle Physics Software Group (BNL)

Brookhaven™
National Laboratory

EIC CompSW Meeting

07/13/2022



Overview

- Data and Analysis Preservation (DAP) is by no means a new topic. Its importance in the past two decades was realized based on prior experience at LEP and later at the LHC – with large amounts of data recorded, and overall complexity of the software required to process it
- Both large RHIC Experiments – STAR and PHENIX – are involved in DAP activities
- This stack of slides is informed primarily by participation of the author in the DAP-related work for PHENIX and also in the DPHEP Collaboration led by CERN
- The goal today is to present some “lessons learned” and see if it can be useful for formulation of the initial EIC DAP policies and practices

DAP Motivations (as they relate to the EIC)

- Traditionally, the goals of DAP are to provide new insights into previously analysed data, continue/finalize some analyses over an extended period of time, test new models, create educational and outreach opportunities etc

- However it is also true that implementing DAP helps ensure reproducibility of the results not only on the timescale typically associated with “preservation” – e.g. decades – but also on a much shorter timescale e.g. publication/conference cycle, for the simple reason that at its core it consists of a number of best practices of scientific computing and knowledge and data management

- cf. – even before an experiment starts, reproducibility of the simulation results needs to be guaranteed, documentation ready and accessible etc

DAP: Challenges and observations

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"

Knowledge Management

- Need to keep records of software provenance, dependencies, configuration, use 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 legacy solution which is a requirement to record such info in a dedicated section of the “Analysis Note” which must accompany every paper, but in reality its efficacy is variable and often insufficient
- Hard to provide continuity of know-how as people move on
- Knowledge dissipates quicker than most people realize

DAP Components

Bit preservation

Analysis capture (Knowledge Management)

Software, Images, CI, Configuration

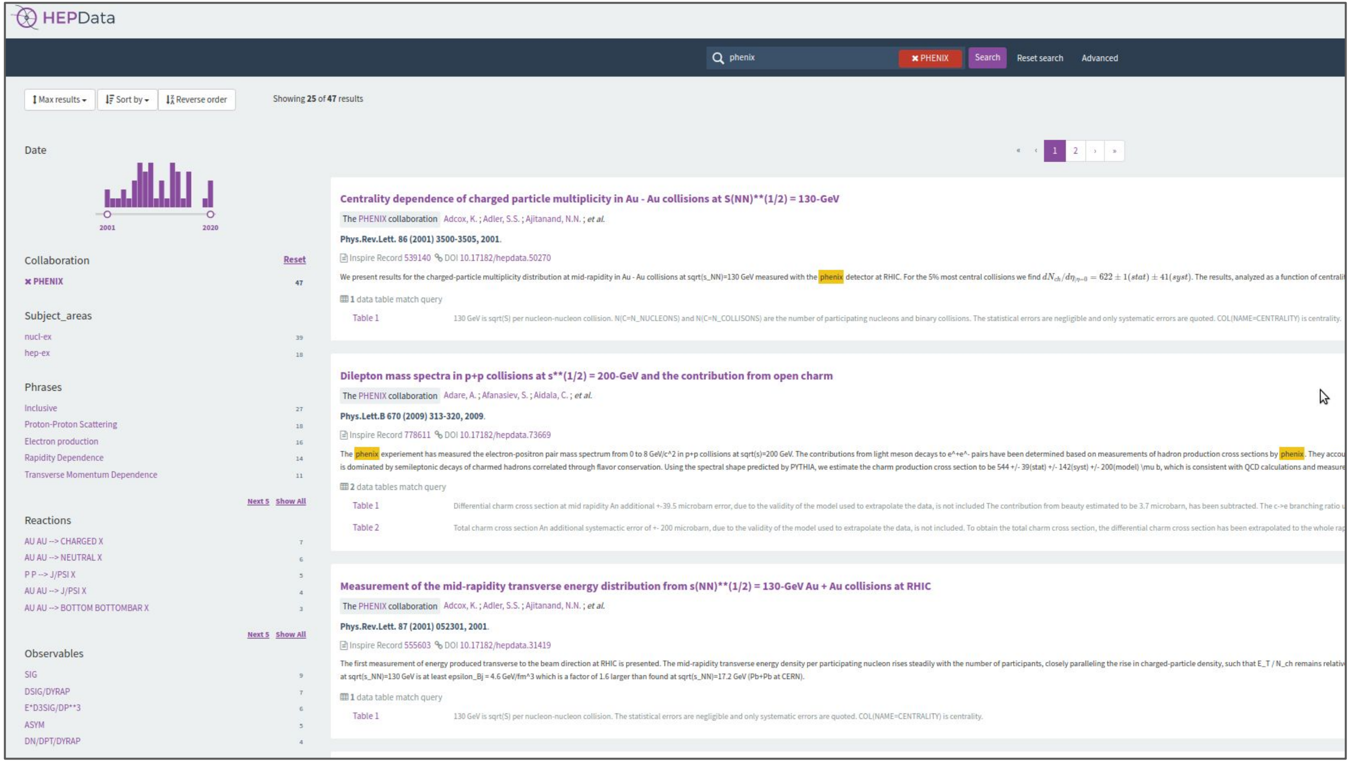
Web-based documentation

Modern repositories and portals for research materials

Tiers of Data Access (incl. HEPData and OpenData)

- Level 1: Data Products used in publications.
  - Such as data points and errors used in plots, in numeric format
  - cf. the “HEPData” portal: https://www.hepdata.net/
- Level 2: Special Purpose Datasets (e.g. 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 (e.g. based on policy)
  - 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.

Experience with HEPData (Tier-1)



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

Date

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 ra

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.

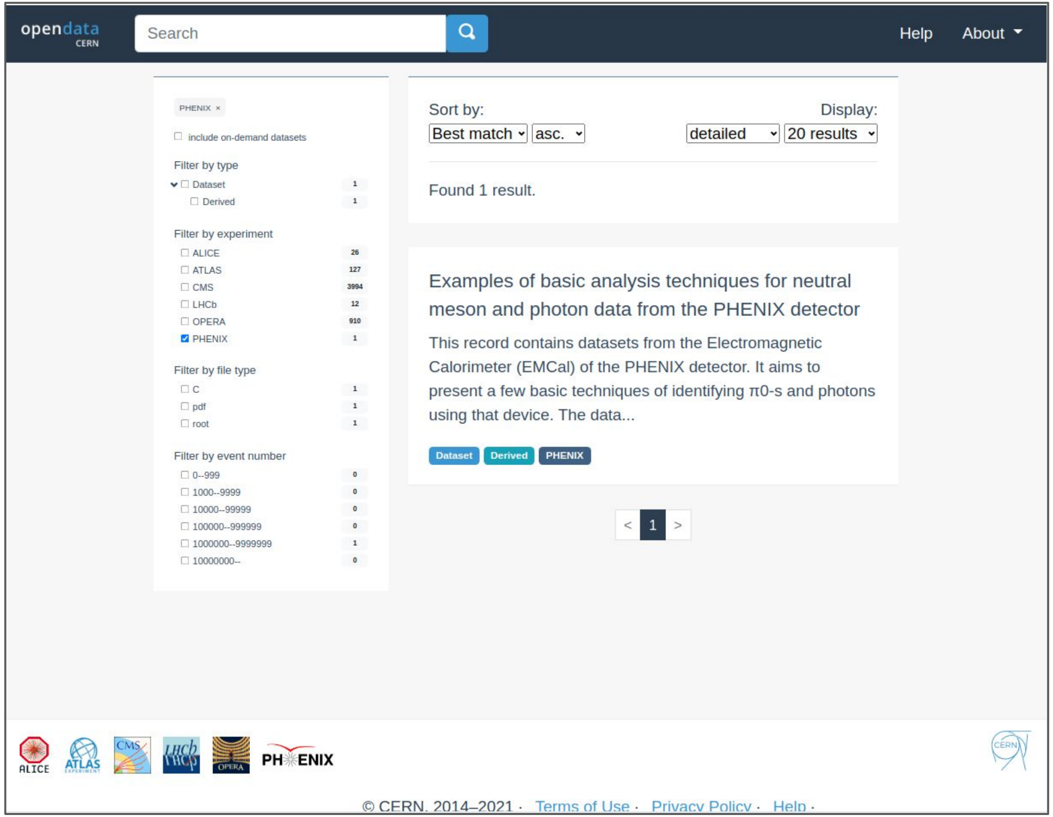
HEPData submissions are a standard practice at the LHC

Now mandated for all new publications in STAR and PHENIX

Revisiting older publication materials as time permits, using GitHub for material development in PHENIX

With established practice and protocols in place, the extra workload is not unreasonable

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



opendata CERN

Search

Help About

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 >

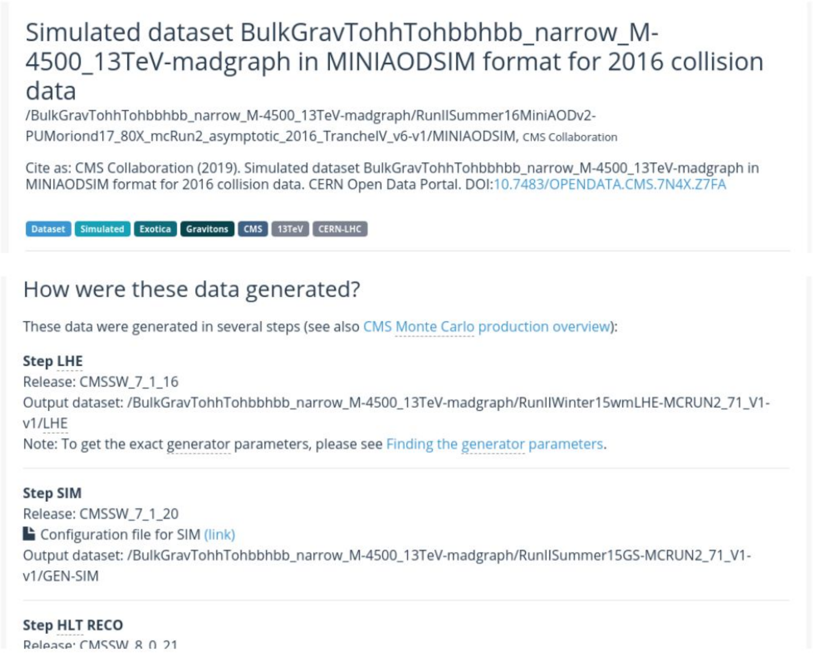
ALICE ATLAS CMS LHCb OPERA PHENIX CERN

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

- Tier-2/3
- The OpenData portal provides a point of synthesis for software, data and documentation
- Contents of this particular package:
  - Derived data (Ntuples)
  - ROOT macros
  - Detailed instructions (PDF)
- Subject area:
  - Analyses based on the EMcal data
- A promising tool to bring new collaborators up to speed, outreach etc
- Select data published by the LHC experiments

OpenData capturing MC data provenance (CMS experiment)

Example 1: Data provenance of simulated datasets



Simulated dataset BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph in MINIAODSIM format for 2016 collision data

/BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM, CMS Collaboration

Cite as: CMS Collaboration (2019). Simulated dataset BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph in MINIAODSIM format for 2016 collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.7N4X.Z7FA

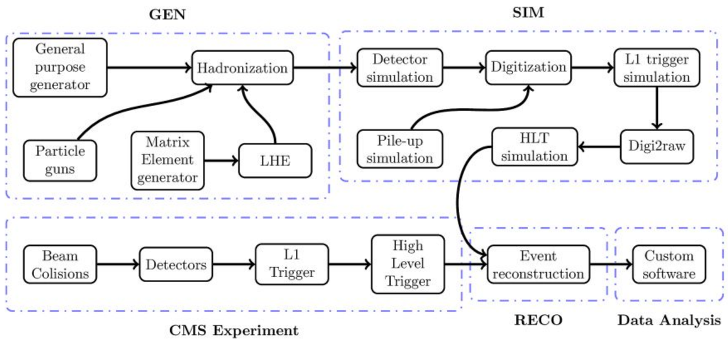
How were these data generated?

These data were generated in several steps (see also CMS Monte Carlo production overview):

Step LHE
Release: CMSSW_7_1_16
Output dataset: /BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIIWinter15wmLHE-MCRUN2_71_V1-v1/LHE
Note: To get the exact generator parameters, please see Finding the generator parameters.

Step SIM
Release: CMSSW_7_1_20
Configuration file for SIM (link)
Output dataset: /BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer15GS-MCRUN2_71_V1-v1/GEN-SIM

Step HLT RECO
Release: CMSSW_8_0_21



- full capture of data generation steps
- full capture of compute environments
- full capture of configuration files
- full capture of production scripts

Data records come with full provenance information

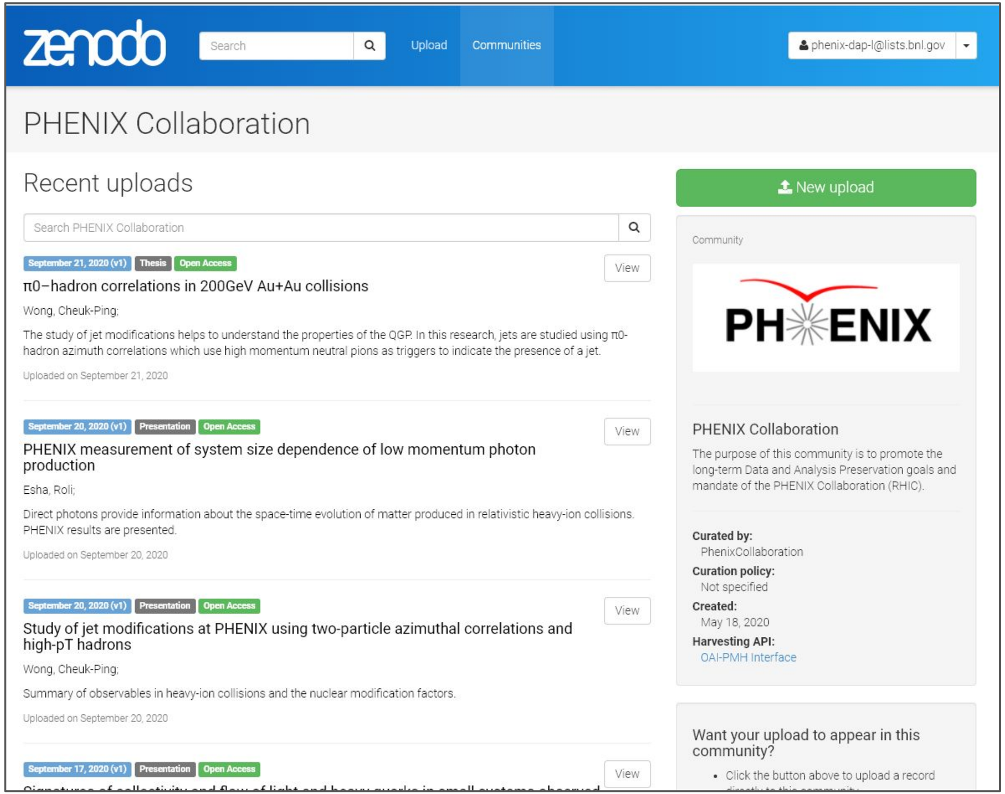
Rivet (for completeness)

- RIVET: Robust Independent Validation of Experiment and Theory
- “Object-oriented C++ framework for analysis algorithms”
- Can import datasets (experimental data) from HEPData
- The interface to MC event generators is HepMC
- …not a priority for EIC at this point but good to keep in mind

Zenodo@CERN – an example of a “community”/catalog

https://zenodo.org/communities/phenixcollaboration

- A modern digital repository
- >500 PHENIX items, uploads ongoing
- Branded, curated, discoverable, DOI
- Well-suited for long-term preservation
  - Also to support current activity: theses, analysis tutorials, presentations etc
  - Can be used to store data in almost any format and aggregation
- Good search capability
  - Keywords (especially if managed)
  - Elastic search
- Search within the “community” helps to improve results (i.e. cut on false positives)



PHENIX Collaboration

Recent uploads

Search PHENIX Collaboration

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

Community

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 directly to this community.

Zenodo – a few notes

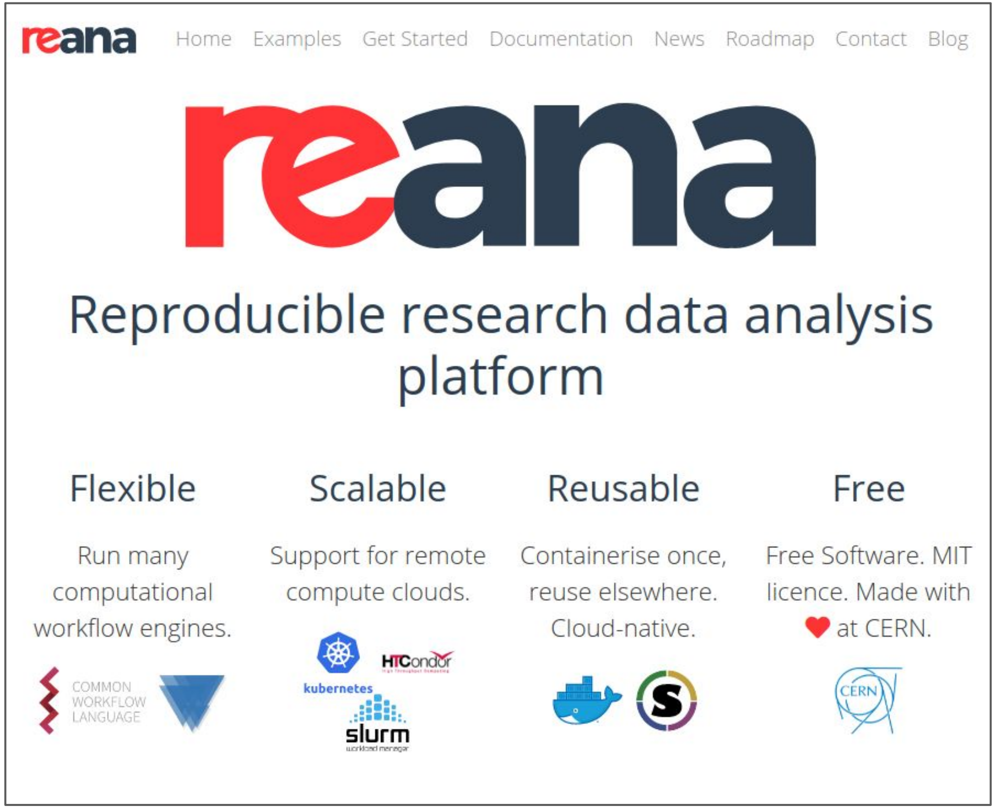
- Zenodo has uses beyond storing just documents, images etc – e.g. it's used in astrophysics to store reference simulation data samples
- Citable software (DOI's issued for GitHub releases)
- A few tiers of access to documents, from public to restricted to private
- It does not handle the document workflow process e.g. comments and approvals, it was not built for that; however it does support versioning
- DOI capability is the key for long term preservation (existence of a migration mechanism is implied/suggested/committed to)
- In my opinion the most viable solution, with main limitation being the available effort for deployment and maintenance – can this be overcome by the EIC community in the long run, given the importance of such system? Can we get more help from CERN?

Capturing the Software Environment

- We use images to capture a few PHENIX SW environments
- Part of the solution for the changing OS’s and dependencies
- NB. interoperability between Docker and Singularity, and the CI angle
- In PHENIX, 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*
- BNL SDCC provides crucial support in the creation of the “main” image
- CVMFS is used to offload large blocks of binary content thus keeping the image size reasonable, for the “maximal” image representing the complete environment
- NB. ideally, containers should be able to run *anywhere*

REANA

- The PHENIX team successfully ran a few simple and one complex analysis cases
- The workflow description syntax in REANA is a clear improvement compared to a free form assembly of shell and other scripts, as it establishes a minimal structure (YAML)
- The learning curve is particularly easy for linear workflows; but DAGs may be complex
- Individual images can be set for steps in workflows i.e. quite a bit of flexibility
- Experience and REANA instances exist at CERN, U Manitoba, BNL etc
- Application for MC seems straightforward



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.

REANA – a few notes

REANA allows the user to record crucial components of analyses:

- The software environment (by reference to images and libraries, environment etc)
- The workflow(s), in one of the available YAML formats
- Data components to be staged in and staged out, any other auxiliary files that are need

The “workspace” paradigm (essentially a sandbox) enforces completeness of the description and provision of well-defined dependencies.

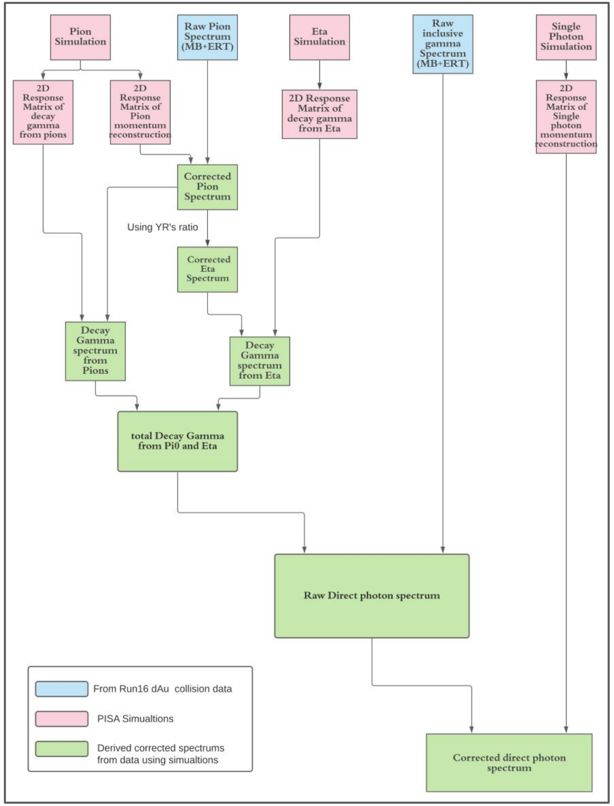
Also of note is a good CLI, a full Python API and Jupyter integration (e.g. one can open a notebook inside a workspace). The workspace is persistent (if needed).

A variety of computational back-ends is supported (even simultaneously – via hybrid pipelines)

Active community, ongoing development and integration

Sidenote: workflow diagrams – a low tech but effective tool

- One the right – the actual flowchart used in adopting the direct photon/EMCal analysis in PHENIX, to REANA
- Complementary to a good verbal description
- Now being considered as a potential requirement for future analysis notes in PHENIX, as an effective and relatively low-cost policy
- In combination with REANA, enhances knowledge sharing and transfer within and between working groups


Pion Simulation
Raw Pion Spectrum (MB+ERT)
Eta Simulation
Raw inclusive gamma Spectrum (MB+ERT)
Single Photon Simulation
2D Response Matrix of decay gamma from pions
2D Response Matrix of Pion momentum reconstruction
2D Response Matrix of decay gamma from Eta
2D Response Matrix of Single photon momentum reconstruction
Corrected Pion Spectrum
Using YR's ratio
Corrected Eta Spectrum
Decay Gamma spectrum from Pions
Decay Gamma spectrum from Eta
total Decay Gamma from Pi0 and Eta
Raw Direct photon spectrum
Corrected direct photon spectrum
From Run16 dAu collision data
PISA Simualtions
Derived corrected spectrums from data using simualtions

A note on “services” (from slide 4)

- There can be a variety of conditions-type data (including calibrations) which needs to be preserved and accessible for sensible DAP – so there must be a plan to keep them up and running over a long period of time, or be able to persist them in storage and spin them up on demand, other options…
  - There is experience at CERN with images containing “embedded” CVMFS
- While small “data artifacts” – custom channel maps etc – can be packaged with the analysis software in images or otherwise, it doesn’t negate the need for a Conditions Database which is operational over a long period of time and is available “anywhere”. Should we store conditions-type derived data used in individual analyses in the Conditions DB? This might help reproducibility since it takes care of version control
- File/data catalog is another service that must be built with durability in mind

Lessons learned

- DAP: *plan and start early*
  - The effort will pay for itself by increasing overall productivity of the experiment
  - Will be hard or impossible to “catch up” later
- Avoid building in-house information systems, there are many tools available
  - State-of-the-art services such as GitHub, Zenodo, OpenData, HEPData, REANA, Rivet, Inspire (publication catalog) etc cover a vast majority of the experiments’ needs
  - There must be no coupling to a particular MC/reco framework
- Containerization solves many of the challenges of capturing the software environment
  - Use it the right (portable) way, with services (DB) made accessible
- Create websites for the long haul (static site generation works well)
  - Avoid platforms that will require updates and maintenance in the long term e.g. Drupal
  - Any resource will become overgrown/obsolete in absence of a designated team of editors
  - Avoid resource fragmentation

Access control on static websites (e.g. Jekyll-based)

- Sites are always deployed on web servers such as Apache or nginx
- Both of these platforms have auth/auth mechanisms that can be leveraged
- A Jekyll site deployed on nginx, with restricted access, has been tested
- Admittedly role-based access is harder to implement

Effort profile?

- Example: a DB expert can create and maintain a database for an experiment etc
- DAP is different – it's a true team effort and commitment (one or two experts cannot achieve the goals e.g. knowledge management) – cf. it is impossible to know enough details of someone else's analysis or MC to effectively reproduce/preserve the study – it has been tried without success
- Of course subject experts will be helpful in any case
- Ideally, working group conveners would agree to a set of policies and practices (e.g. designate a few critical simulations for REANA and/or OpenData deployment)
- Historically, DAP has been an underfunded area at RHIC, while some LHC experiments allocated a healthy level of effort to DAP
- There is a fraction of FTE available to help now (yours truly)

DAP@CERN

- CERN is the preeminent center of DAP expertise and a host to a number of world-class DAP facilities
- The DPHEP Collaboration is the hub of this activity https://dphep.web.cern.ch/
- Example: BNL is a member of the DPHEP Collaboration, both the facility and individual experiments (PHENIX) participated in the DPHEP meetings and other activities
- It would make a lot of sense for the EIC Project Detector collaboration to seek association with DPHEP, to leverage the considerable expertise and resources available through it

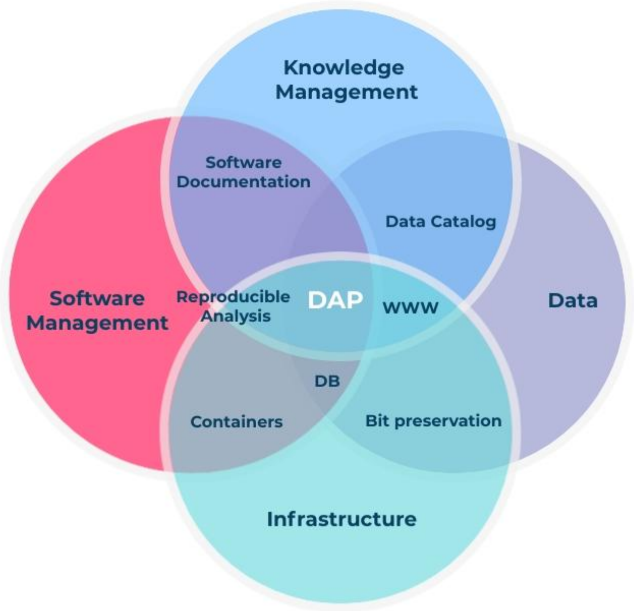
Final Note

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

DAP focus on knowledge management is conducive to efficient knowledge transfer within the collaboration and across projects (cf. onboarding graduate students)

Software management, packaging and containerization facilitates deployment

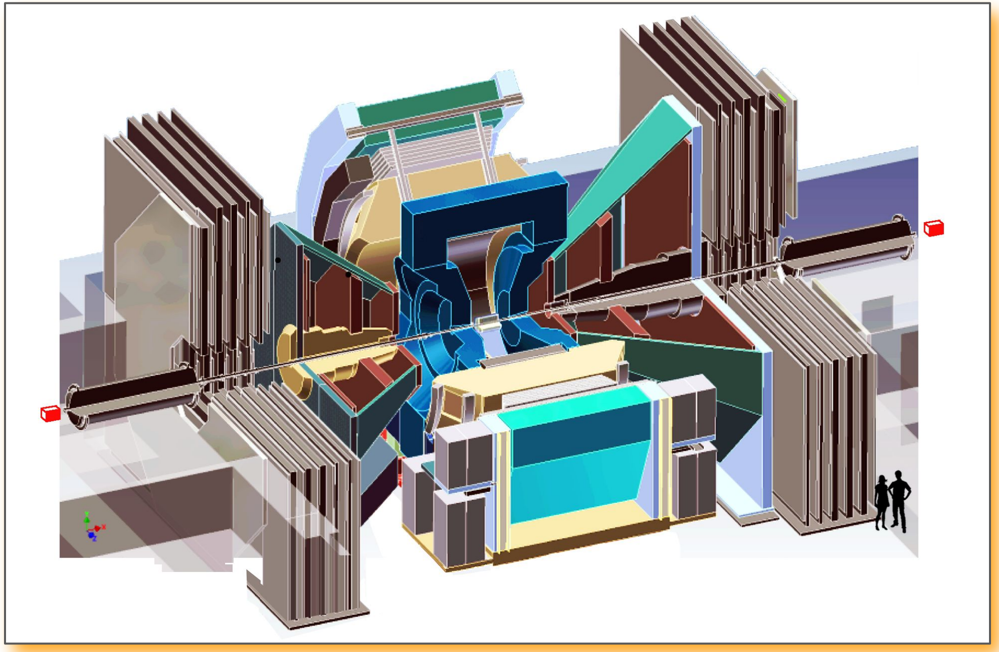
Modern digital repositories create efficient document management solutions on any time scale (cf. OpenData, Zenodo etc)



Backup

PHENIX

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

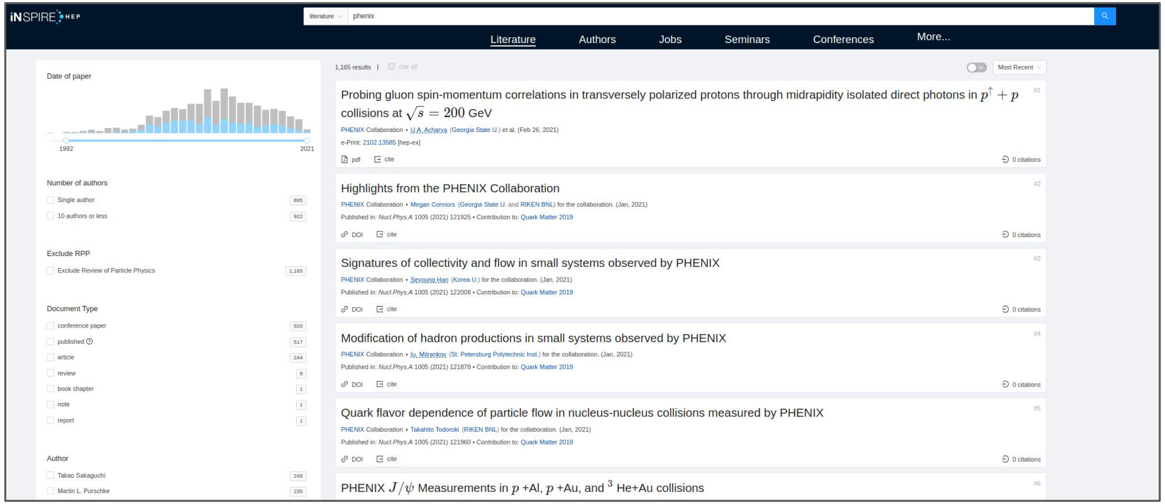


- For more details please see the “PHENIX Collaboration Community” on Zenodo: https://zenodo.org/communities/phenixcollaboration/

PHENIX today

- Data taking finished in 2016 with ~24PB of raw data accumulated
- Active analysis work underway (average ~10 articles a year in 2019-2022)
  - Total of >240 published papers + conference contributions (total ~1300 items), ~170 PhD theses and counting

- All current analyses are done using preserved data



Date of paper

Number of authors

□ Single author 895
□ 10 authors or less 922

Exclude RPP

□ Exclude Review of Particle Physics 1,165

Document Type

□ conference paper 920
□ published 517
□ article 244
□ review 9
□ book chapter 1
□ note 1
□ report 1

Author

□ Takao Sakaguchi 249
□ Martin L. Purschke 235

1,165 results

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

PHENIX Collaboration • U.A. Acharya (Georgia State U.) et al. (Feb 26, 2021)

e-Print: 2102.13585 [hep-ex]

0 citations

#2 Highlights from the PHENIX Collaboration

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

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

0 citations

#3 Signatures of collectivity and flow in small systems observed by PHENIX

PHENIX Collaboration • Seyoung Han (Korea U.) for the collaboration. (Jan, 2021)

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

0 citations

#4 Modification of hadron productions in small systems observed by PHENIX

PHENIX Collaboration • Iu. Mitrankov (St. Petersburg Polytechnic Inst.) for the collaboration. (Jan, 2021)

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

0 citations

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

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

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

0 citations

#6 PHENIX J/ψ Measurements in p +Al, p +Au, and ^{3} He+Au collisions

PHENIX: Legacy web infrastructure

- Information was spread across a few legacy web resources - the software, detector and subsystem information and other documentation

- Information was diluted with items once relevant for PHENIX but no longer current or aligned with its current and future needs

- PHP-based proprietary information systems e.g. document database (papers, talks, theses), numerical data archive etc became difficult to upgrade, maintain and keep secure - and experiences outages

PHENIX: Software challenges

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