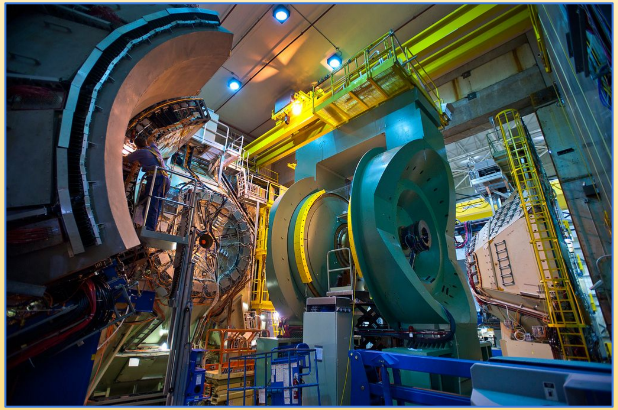
Current status of the Data and Analysis Preservation effort in the PHENIX experiment

Maxim Potekhin
Nuclear and Particle Physics Software Group
Brookhaven National Laboratory

PV2023

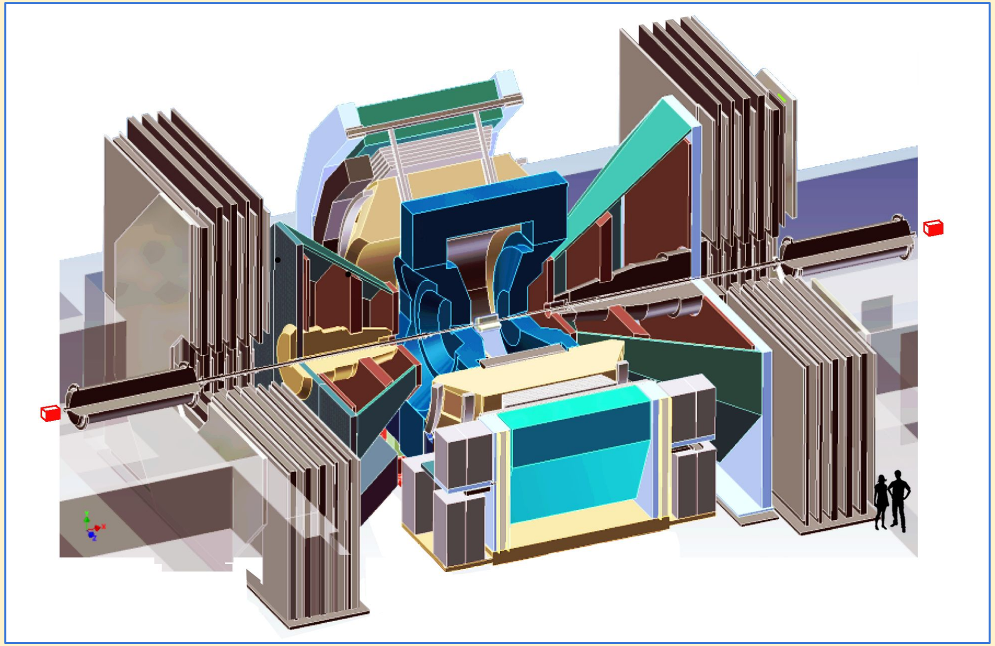


May 3, 2023

Overview

- PHENIX in a nutshell
- Data and Analysis Preservation (DAP) challenges in PHENIX – and technical solutions identified, leveraged or developed to meet its goals
- Description of an effort currently in progress, to preserve the direct photon analysis based on the PHENIX data

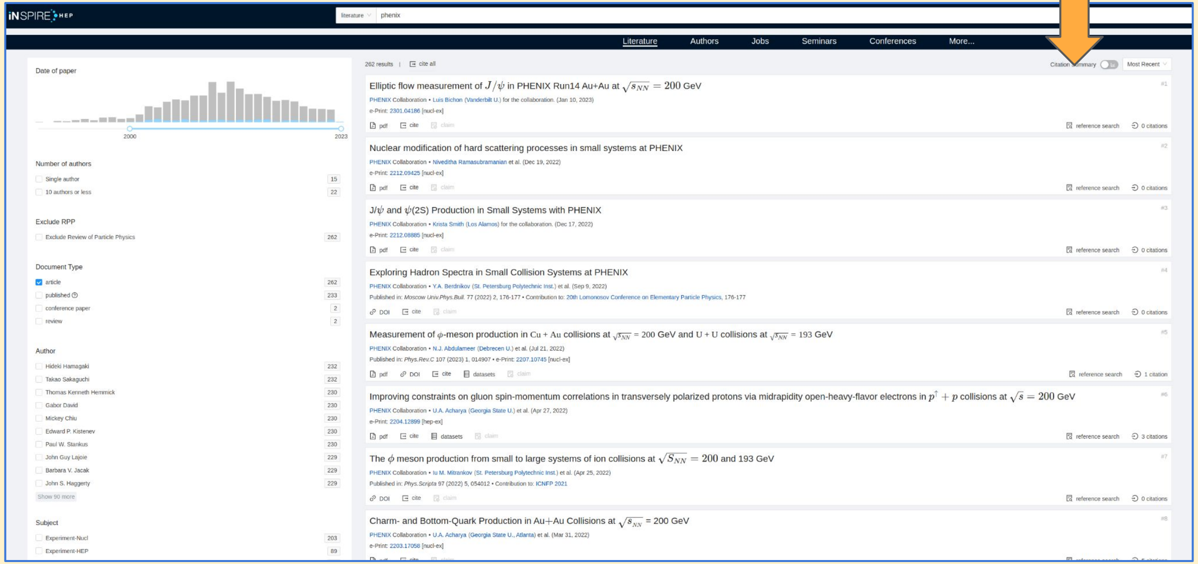
PHENIX in a nutshell



- "Pioneering High Energy Nuclear Interaction eXperiment"
- A complex general purpose detector optimized for high rates
- Search for QGP and study of its properties, spin physics, electromagnetic probes, dimuons
- 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 approx. 24PB of raw data accumulated
- Active analysis work underway (average ~10 articles a year in recent years)
  - About 260 published papers + *many* conference contributions, a total of >1200 entries on InspireHEP
  - Approx. 160 PhD theses



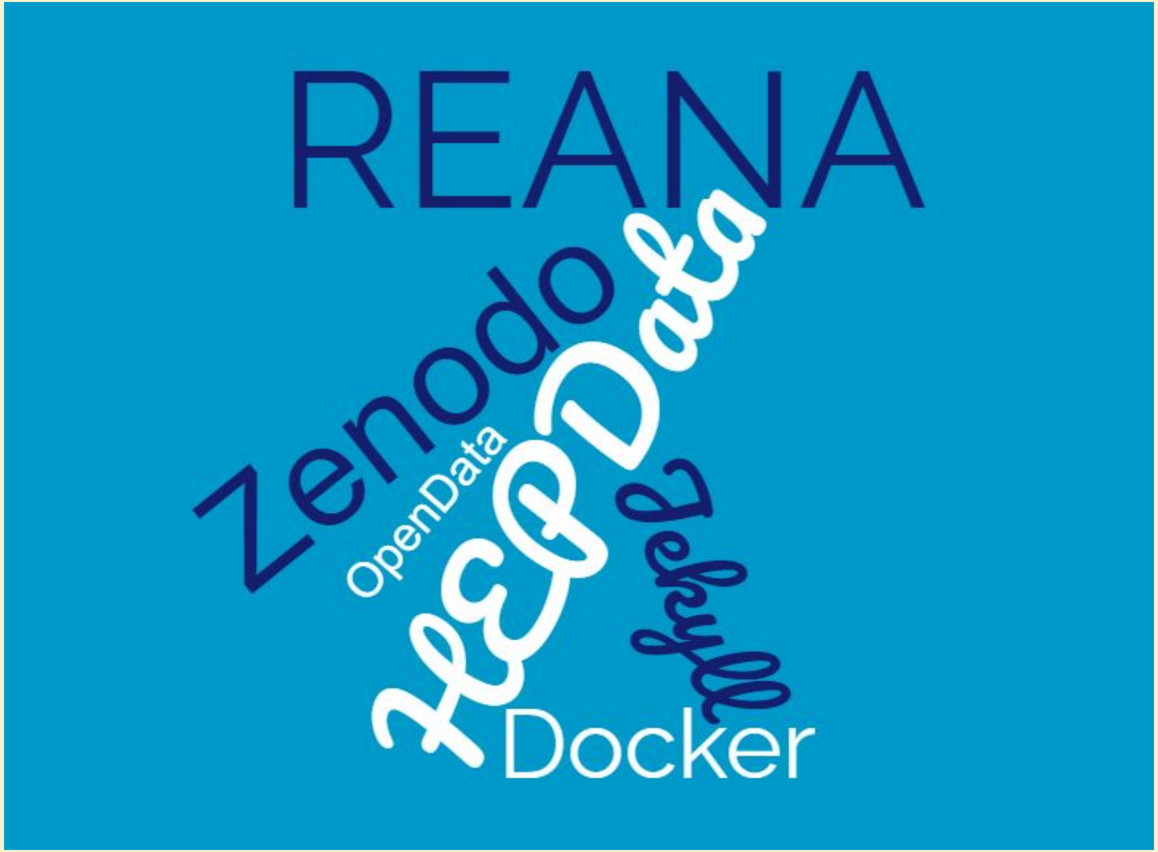
- Key active analyses
  - Heavy flavor in central and forward rapidity
  - Low p_T photons and thermal dileptons
  - High p_T direct photon and photon-jets

Motivations for DAP – data and analysis preservation

- In general, the goal of DAP is to have a reproducible analysis capability over a long period of time so as to retain the value of the data and effort
- However, it also brings benefits on the shorter time scale
  - Reliable reproducibility is a necessary component of any analysis, and DAP tools are well suited for that
  - It aids the capability to perform a modified or new analysis within the same framework
  - Onboarding new researchers is facilitated by knowledge management inherent in DAP, as well as good software management practices that it engenders
- The policy factor: funding agencies increasingly require both new and existing experiments to develop and implement plans for Open Data, which in practice implies not just preserving the data but also the tools and documentation necessary to access it.

Challenges

- Web infrastructure
  - Without constant effort the web resources become obsolete, fragmented and the underlying technology (e.g. legacy versions of PHP, databases) can become difficult to maintain, especially with limited resources. Network security presents its own set of issues.
- Knowledge management
  - As people move on to other projects, continuity of know-how becomes an issue. Software, detector and other documentation must remain accessible and useful in the long term.
  - In the past, research document management was done using in-house solutions which share common problems with the legacy web infrastructure
- Software
  - OS versions, compilers, tools and components are not constant in the long term. Analysis preservation implies that the exact software configuration and services are captured and remain operational


REANA
Zenodo
HEPData
OpenData
Jekyll
Docker

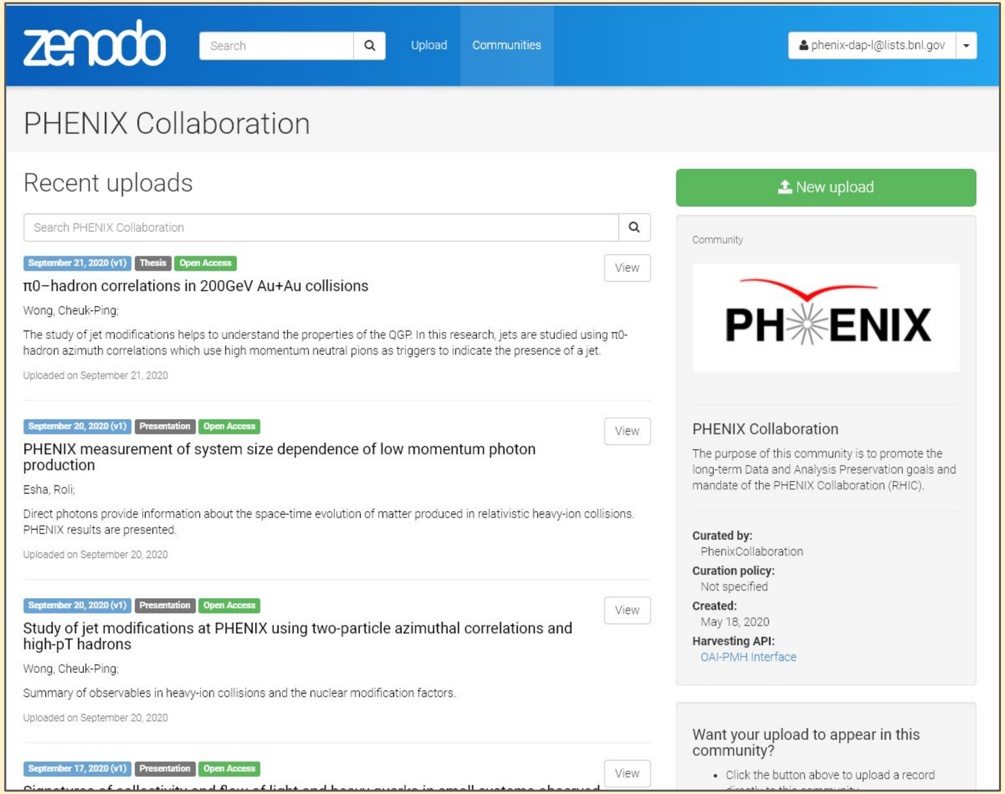
Solutions

- Web infrastructure *(aging, fragmented, hard to maintain)*
  - Redesign of the PHENIX website with long-term maintenance in mind, utilizing a popular static website generation technology "Jekyll": phenix.bnl.gov
  - Consolidation, curation and reformatting of materials from previous web resources
  - Vastly improved security due to static nature of the site (important)
- Knowledge management *(hard-to-discover, custom information resources)*
  - Leveraging modern repositories and web portals to host research materials, replacing previous in-house solutions
  - A "Controlled Vocabulary" of keywords to make materials discoverable, e.g. conferences with PHENIX participation, physics topics/terms, detector components, software technologies – complements sophisticated Zenodo search capabilities
- Software *(evolving platforms, compilers, components)*
  - Containers (Docker+Singularity): capture of the software environment, utilization in REANA

Zenodo@CERN – the PHENIX page

https://zenodo.org/communities/phenixcollaboration

- A modern digital repository
- >600 PHENIX items, uploads ongoing, including presentations from >100 conferences and 160 PhD theses
- Branded, curated, findable, with DOIs
- Well-suited for long-term preservation
  - …and also current activity: theses, analysis tutorials, presentations etc
  - Can be used to store data in almost any format and aggregation
- Excellent search capability
  - Keywords (including “controlled”)
  - Elastic search



PHENIX keywords on the website ("vocabulary")

Experiment | Results | Detectors | Offline Software | Analysis

Physics (97 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 c
cgc	Color Glass Condensate (type o
charm	Particles containing the c-quark
charmonium	Meson containing a c-quark an
cnm effects	Cold Nuclear Matter effects
correlations	Various types of correlations
cronin effect	Cronin effect
cross section	Cross section (as it applies to sc
cu+au	Copper-on-gold collisions
cu+cu	Copper-on-copper collisions
cumulant	Cumulant
d+au	Deutron-on-gold collisions
d-meson	D meson
dca	Distance of Closest Approach
dielectron	A pair of electrons
dilepton	A pair of leptons
dimuon	A pair of muons produced in a
direct photon	Direct photons produced in nuc
drell-yan	Drell-Yan type of process
electron	Electron

Experiment | Results | Detectors | Offline Software

Conferences (97 items)

Keyword	Description
aum16	RHIC & AGS Annual Users Meeting (2016)
aum17	RHIC & AGS Annual Users Meeting (2017)
aum18	RHIC & AGS Annual Users Meeting (2018)
aum19	RHIC & AGS Annual Users Meeting (2019)
aum20	RHIC & AGS Annual Users Meeting (2020)
aum21	RHIC & AGS Annual Users Meeting (2021)
aum22	RHIC & AGS Annual Users Meeting (2021)
charm21	10th International Workshop on CHARM Physics
cipanp18	Conf. on the Intersections of Particle And Nuclear Physics (2018)
cipanp22	Conf. on the Intersections of Particle And Nuclear Physics (2022)
cpod17	Critical Point and Onset of Deconfinement (2017)
cpod18	Critical point and onset of deconfinement (2018)
dis17	Deep Inelastic Scattering (2017)
dis18	Deep Inelastic Scattering (2018)
dis19	Deep Inelastic Scattering (2019)
dis21	Deep Inelastic Scattering (2021)
dis22	Deep Inelastic Scattering (2022)
dnp19	DNP (2019)
dnp20	DNP (2020)
epshep17	EPS HEP 2017
eunpc22	European Nuclear Physics Conference 2022
fwph21	Workshop on forward physics and QCD (2021)
ghp17	7th Workshop of the APS Topical Group on Hadronic Physics (2017)
ghp19	8th Workshop of the APS Topical Group on Hadronic Physics (2019)
hfwinc22	Heavy Flavour Production in Nuclear Collisions (2022)
hp16	Hard Probes 2016
hp18	Hard Probes 2018
hp20	Hard Probes 2020
hptlhc19	High pT physics in the RHIC/LHC era (2019)
hq18	Hot Quarks 2018

Each keyword is a link to a functioning Zenodo query

Example of a subsystem page on the PHENIX Website

Experiment | Results | Detectors | Offline Software | Analysis

Electromagnetic Calorimeter

Write-ups

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

Theses

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

Publications

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

Presentations

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

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

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

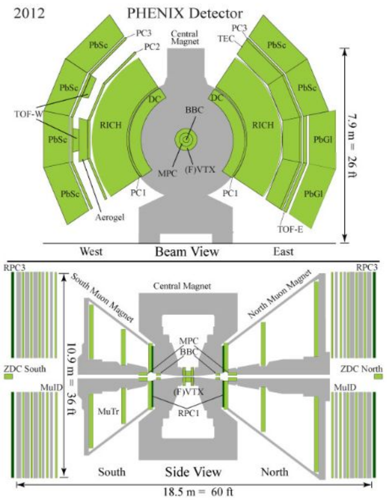
Run summary pages (rebuilt on the new site)

Experiment | Results | Detectors | Offline Software | Analysis | About

01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16

Run 12

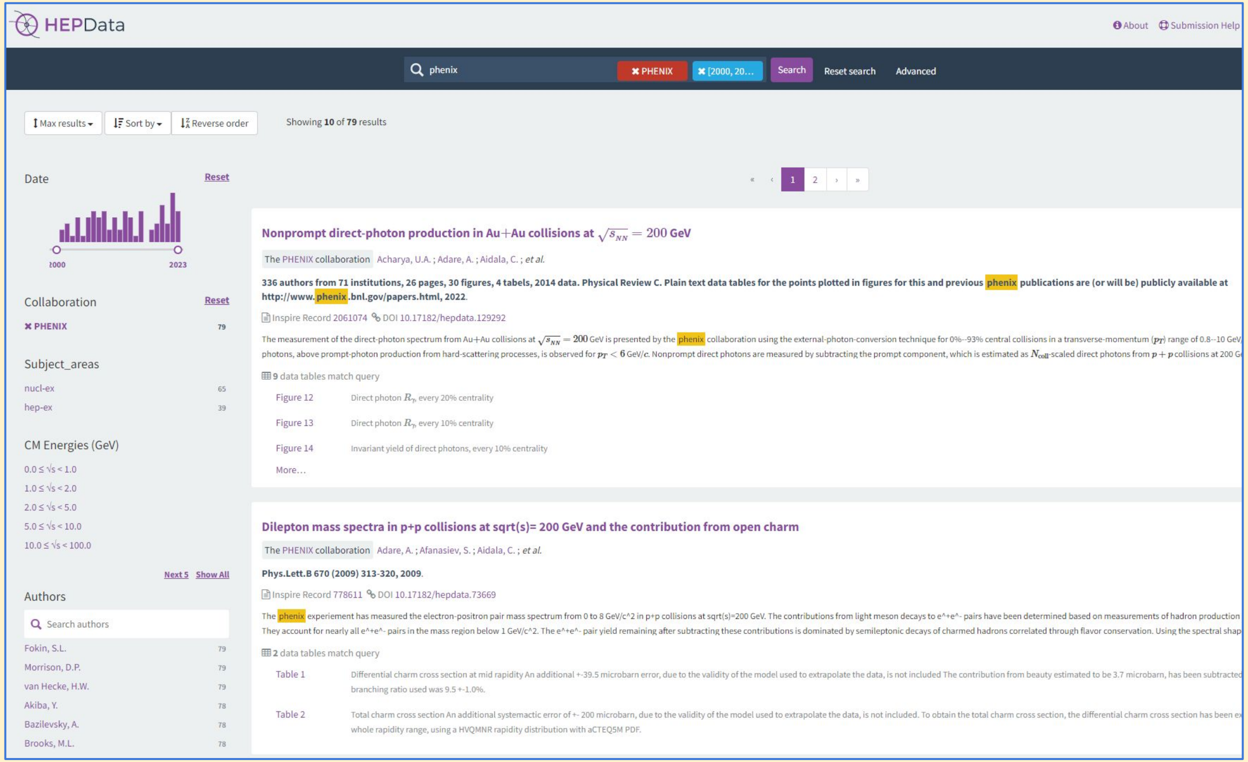
Configuration Diagram



RHIC+PHENIX Run Records

Species	Energy (GeV/nucleon)	Integrated Luminosity [Polarization L/T]	N_{events} [BBC_{30cm}/BBC_{narrow}]
polarized p+p	100.2	– /10 pb^{-1}	
polarized p+p	254.9	32/– pb^{-1}	
$^{238}U^{92+}+^{238}U^{92+}$	96.4	0.2nb^{-1}	1.2B/0.8B
$^{63}Cu^{29+}+^{197}Au^{79+}$	99.9+100.0	5nb^{-1}	0.8B/8.1B
$^{197}Au^{79+}+^{197}Au^{79+}$	2.5	-	Very short

PHENIX HEPData presence

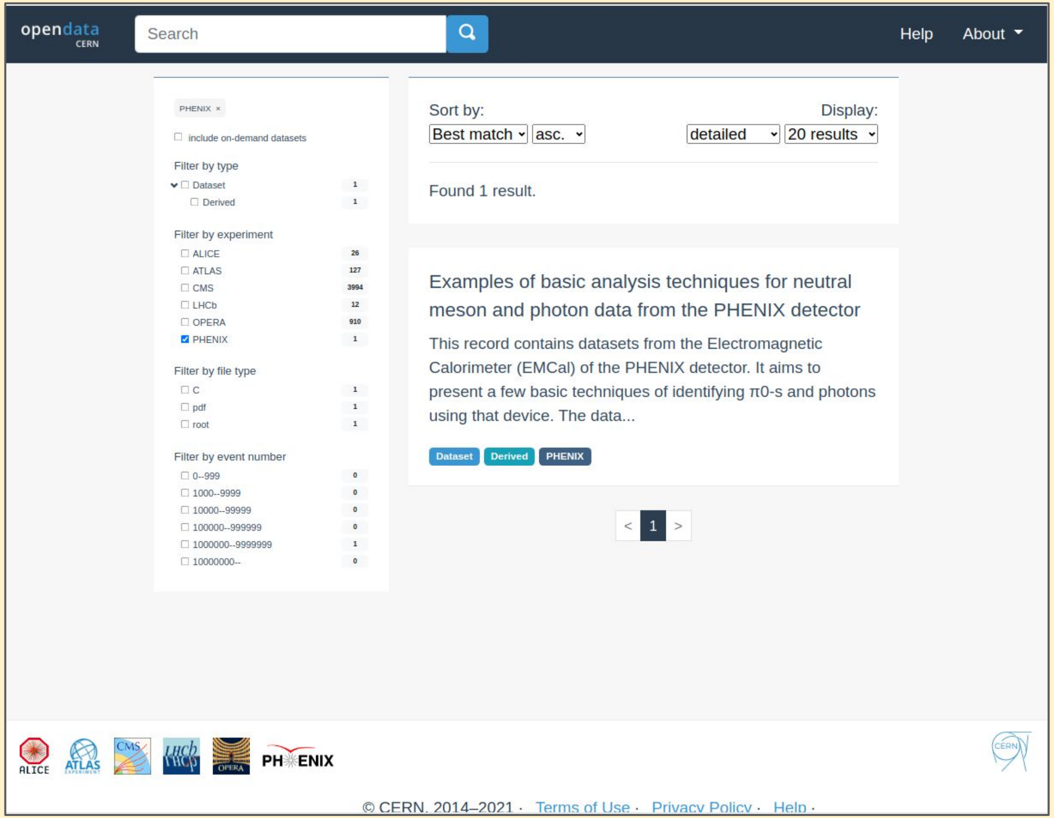


HEPData submissions mandated for all new publications in PHENIX

We are also revisiting older publications, using GitHub to develop materials and coordinate teamwork (with PRs etc)

Within the last 3 years, the number of published PHENIX HEPData entries increased from 23 to 80 – special thanks to the team at UTK

The PHENIX OpenData entry



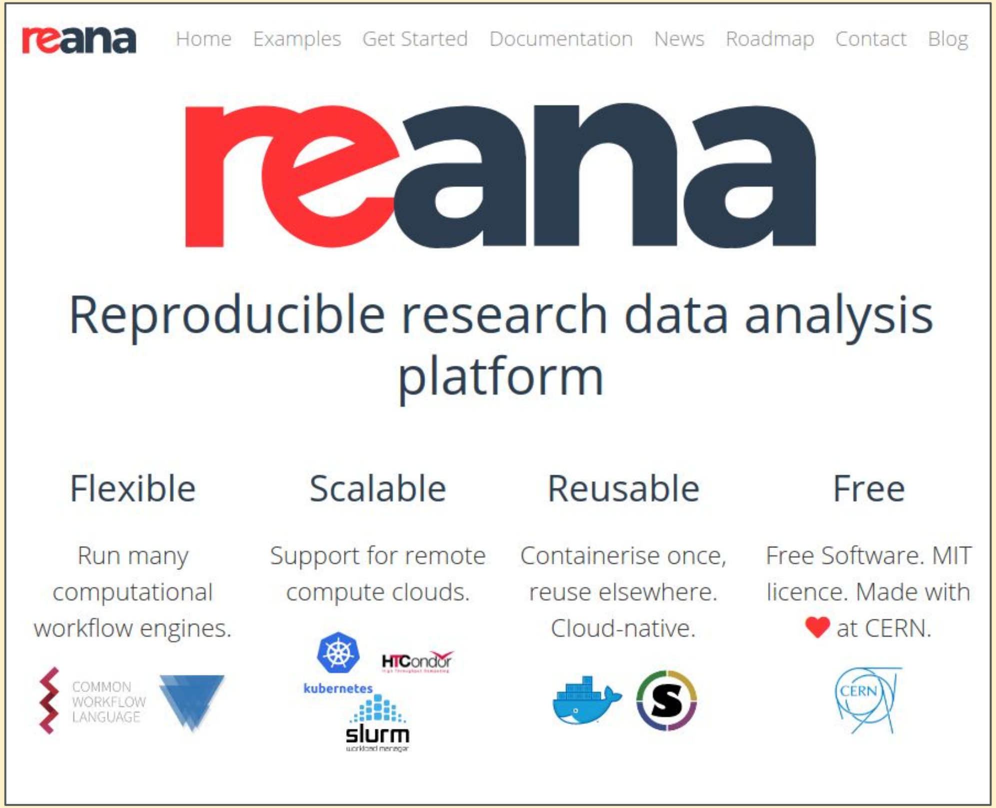
- It's a start…
- OpenData: a point of synthesis for software, data and documentation, capable of handling complex analysis cases and making substantial amounts of data accessible
- Contents of this particular package:
  - Derived data (Ntuples)
  - ROOT macros
  - Detailed instructions (PDF)
- Subject area:
  - Analyses based on the EMcal data

Capturing the Software Environment

- We use images to capture a few PHENIX SW environments, as a solution for the changing OS landscape, software components and dependencies
- Images created for PHENIX range from simple ones, capturing legacy ROOT and compilers, to the complete software stack as it is installed on the facility (large!)
- NB. interoperability between Docker and Singularity, i.e. containers can run in batch
- 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*
- Full images are stored in a private registry and accessible at the facility

REANA

- The core goal of REANA is reproducible analysis
- 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 structured approach to analysis description (YAML)
- The learning curve in REANA is particularly easy for linear workflows; general DAGs are (a lot more) complex
- Individual software images can be set for steps in workflows – more flexibility
- The PHENIX team is currently focusing on one specific REANA analysis – direct photon production in *d+Au* collisions

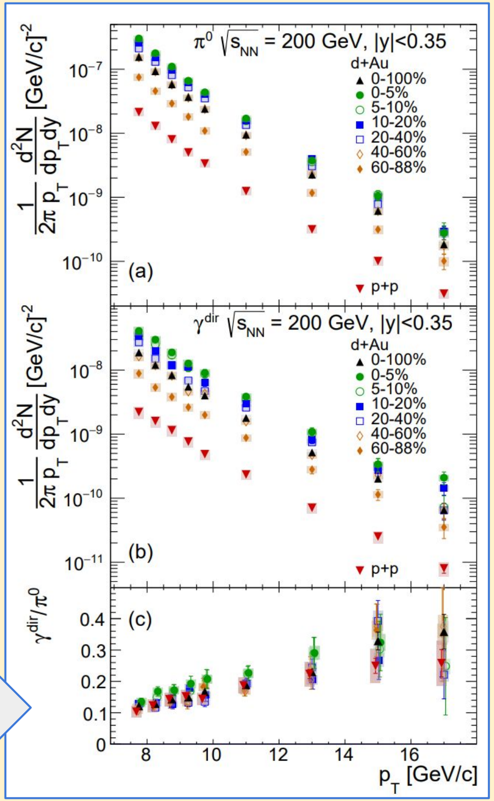


The study – use case

- PHENIX is a relativistic heavy-ion experiment and its main focus is determination of the properties of the Quark-Gluon Plasma (QGP).
- One of the most important signatures of formation of QGP was the jet suppression in heavy ion collisions, quantified by the so-called nuclear modification factor: $R_{AB}(p_T) = Y_{AB}(p_T)/(N_{coll} \times Y_{pp}(p_T))$, where
  - A, B denote the two colliding nuclei
  - Y_{AB} is the inv. yield measured in A+B collisions, and Y_{pp} is the yield measure in pp collisions
  - N_{coll} is the number of binary nucleon-nucleon collisions; its estimate is typically model-dependent (e.g. using the Glauber model)
- The main motivation for this study was to determine whether there can be a bias in centrality estimates in “small systems” such as *d+Au* collisions, resulting in hard-to-explain behavior of the nuclear modification factors. For background and details please see https://arxiv.org/abs/2303.12899

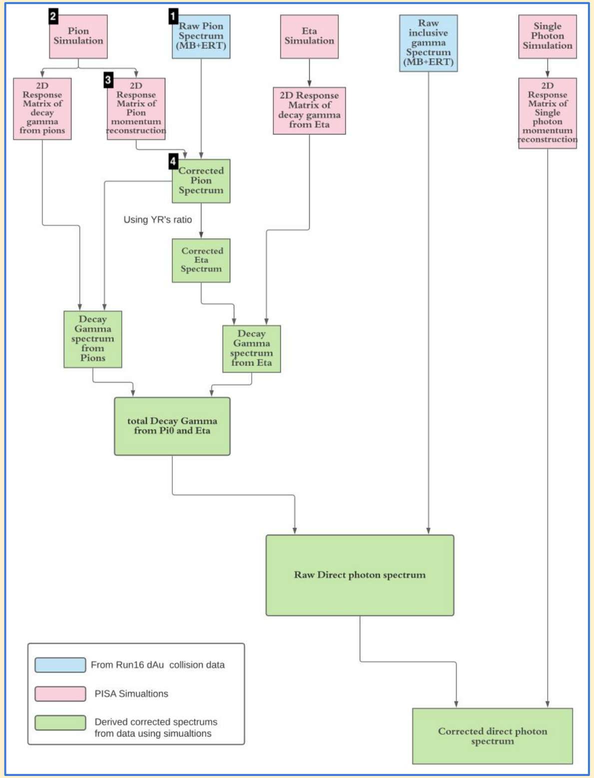
The measured yields of γ^{direct} and π^0, and their ratio: γ^{direct}/π^0

- *d+Au* collisions
- https://arxiv.org/abs/2303.12899
- On the bottom chart: note the similar ratio for most collisions except most central ones
- An important result that warrants analysis preservation
- The Electromagnetic Calorimeter was the principal detector used in this analysis



Workflow diagrams – a low tech but effective tool

- On the right – the actual flowchart used in adopting the direct photon/EMCal analysis in PHENIX, to REANA
- Complementary to a good verbal description, published at, and cross referenced with the textual description of the analysis on the PHENIX website
- Diagrams like this one are 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



The documentation page for this analysis (excerpt)

Experiment | Results | Detectors | Offline Software | Analysis

```
.L Pi0EmbedFiles.C
Pi0EmbedFiles t
t.Loop()
```

In the REANA script, this is used as follows: `cat pi0run.script | root -b`. Note that a PHENIX-specific ROOT library `libTHmul.so` is loaded in the beginning, as this is necessary for proper operation of the macro.

Please refer to the relevant folder in the PHENIX GitHub repository for access to the actual material.

This is the driver script `Pi0EmbedFiles.csh`. Note that symbolic links are created to feed successive files from a holding folder, to the *ROOT* macro.

```
#!/bin/tcsh
source ./setup_env.csh

foreach i (`seq 0 1 $1`)
    ln -s gpfs/mnt/gpfs02/phenix/data_preservation/phnxreco/emcal/Pi0/test/simPi0_$i.root pi0_dAuMB.root
    echo File: $i
    ls -l pi0_dAuMB.root
    cat pi0run.script | root -b
    mv EmbedPi0dAu.root EmbedPi0dAu_$i.root
    rm pi0_dAuMB.root
end
tar -cf embedPi0dAu.tar EmbedPi0dAu_*
```

Processing of input files takes place sequentially and in this case takes a significant amount of time compared to other steps, i.e. a feew hours.

The results of all emedding runs are bundled together in a *tar* archive to make downloading easier. Upon retrieval the data need to be merged using the utility `haddPhenix` which is done in **Block 3** (see below). Upon completion of this step the file `embedPi0dAu.tar` needs to be downloaded, and put in the folder from where the next step is launched. An example of the cownload command, assuming the workflow was named "embed":

```
reana-client download -w embed embedPi0dAu.tar
```

2D Response Matrix of Pion Momentum Reconstruction (Block 3)

The original macro `generationRM_Pi0.cc` was cleaned up (including removal of interactive graphics) and renamed `generationRM_Pi0.C`.

Tar file containing multiple ROOT files (see **Block 2** description above) is uploaded as input for this step. Abbreviated contents of driver script look as follows:

```
#!/bin/tcsh
source ./setup_env.csh
haddPhenix EmbedPi0dAu.root EmbedPi0dAu_*
root -l -b -q 'generationRM_Pi0.C'
```

The macro generates the file `EmbedPi0dAu.root` by merging inputs via `haddPhenix` and produces `Pion_RM.root`. The complete description is in `generationRM_Pi0.yaml`, which resides with all subsidiary scripts in the folder `generationRM`.

The workflow description is as follows:

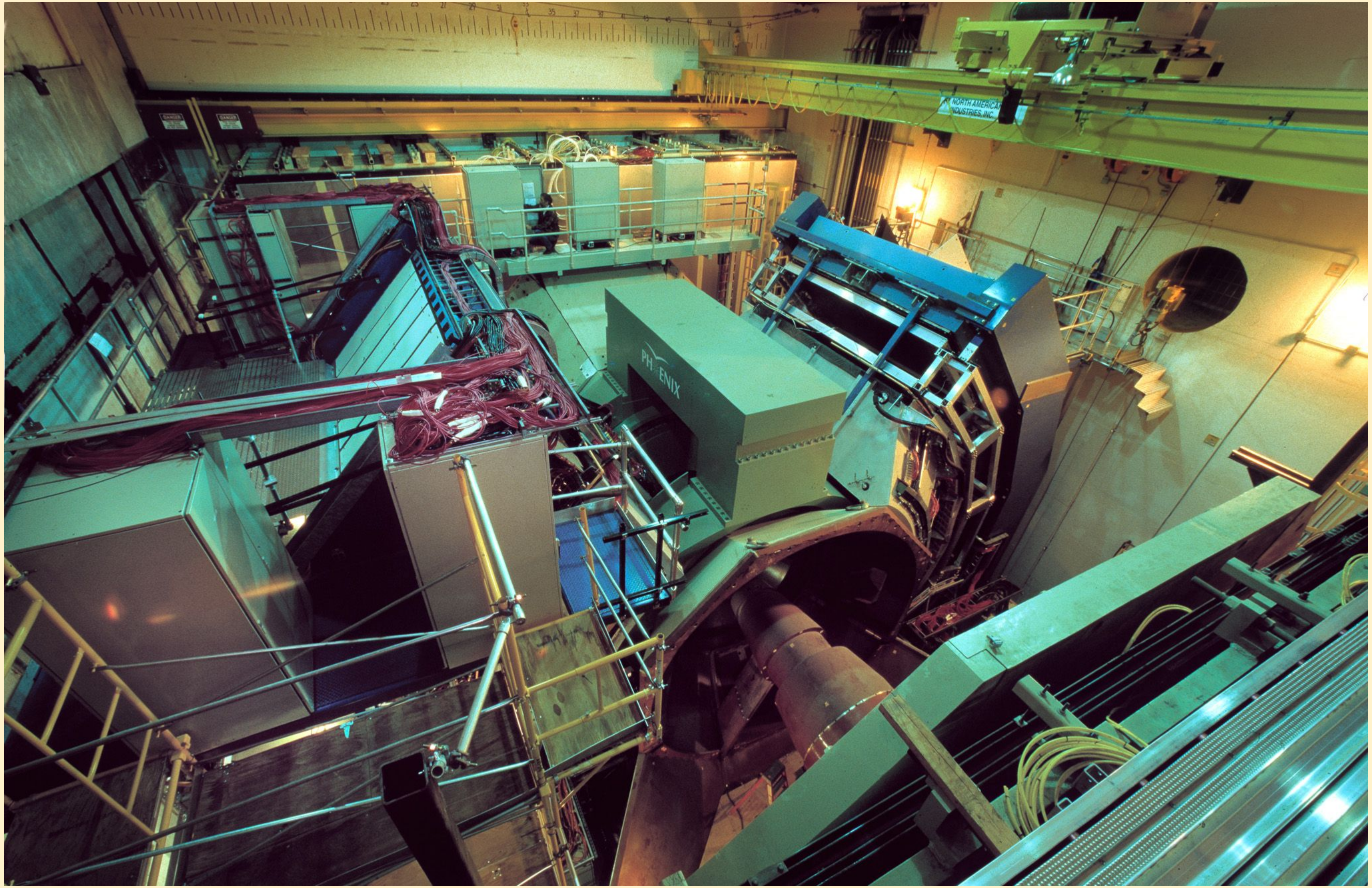
Current status of this study

- The results are public at https://arxiv.org/abs/2303.12899
- The initial version of this analysis was ported to REANA in 2021, it included a set of the most important analysis blocks shown above, but not quite complete
- It is clearly beneficial to expand the part of the analysis converted for reuse and preservation – this is the focus of the current work
- The analysis itself has progressed and improved in the past two year, so this must be reflected in the REANA materials
- This round of updates is expected to be completed in Summer of 2023

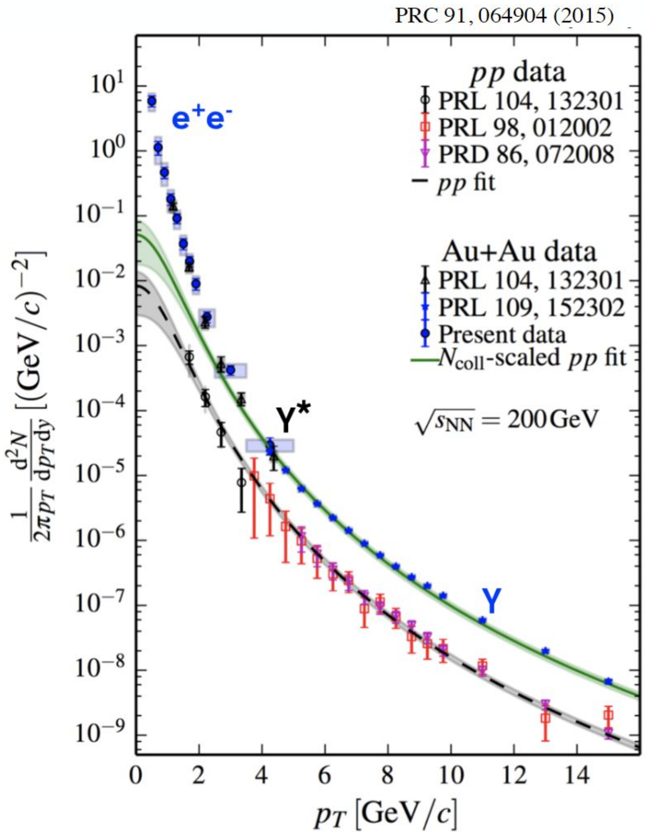
Challenges, Lessons and Plans

- In a legacy experiment
  - Too many components and dependencies may have accumulated over time, to have an unencumbered build on top of a “clean” OS
  - Services (like databases) and conditions data can also be tightly coupled to a particular facility
  - All these factors impede making the software images public
  - Analysis code is rarely documented in sufficient detail, to enable straightforward preservation
- Leveraging existing, community-supported platform allows to make substantial progress in the area of analysis preservation even with limited resources, the support from the facility is instrumental
- “Plan and execute early” is the best advice that new experiments can take
- PHENIX plans to continue its vigorous work of HEPData materials preparation, research document management in Zenodo and analysis preservation on REANA

Backup slides


PHENIX

Direct photon yields for p+p collisions at 200 GeV are consistent with pQCD calculations



Thermal photon yield = Direct photon yield of Au+Au — Hard scattering contribution (N_{coll}-scaled p+p)

The Au+Au yield is consistent with N_{coll}-scaled p+p yield above 4 GeV

Credits:
Roli Esha
Initial Stages 2021

Knowledge Management

- Need to keep records of software provenance, dependencies, configuration, use etc – the “know how”
- 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

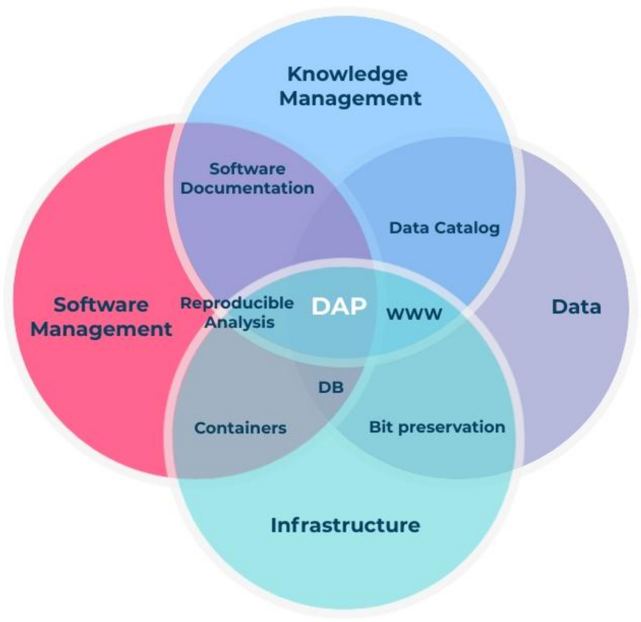
Benefits of DAP

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)



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 editors
  - Avoid resource fragmentation

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.

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)

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"