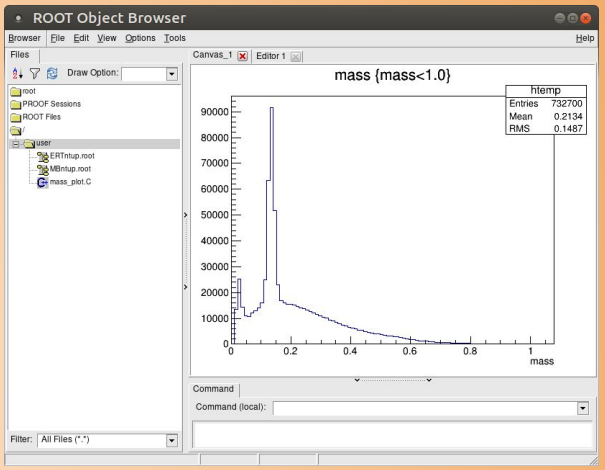
Data & Analysis Preservation: status update



Maxim Potekhin
Nuclear and Particle Physics Software Group

BROOKHAVEN NATIONAL LABORATORY

PHENIX

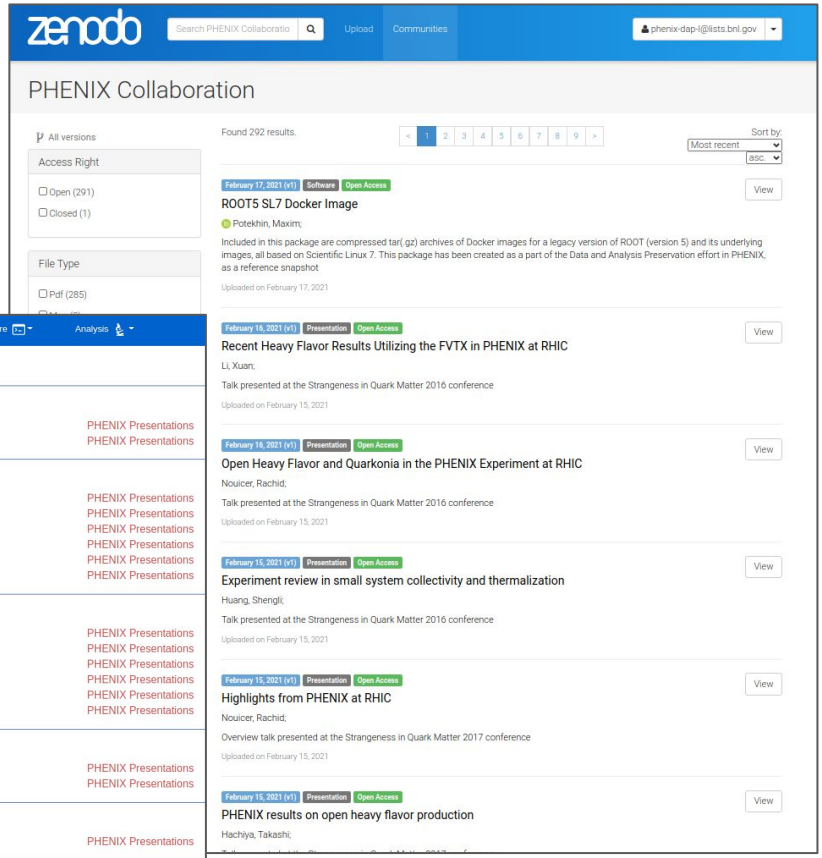
PHENIX Conveners Meeting
03/03/2021

Overview

- Please see slides from my previous presentation in January 2021 for a broad overview of Data and Analysis Preservation in PHENIX
- https://docs.google.com/presentation/d/19ksQ05Y-pU3SuLxIciZsdydjgKJvHlh_UbwtXiQctn4/edit?usp=sharing
- Topics for today
  - Zenodo
  - Progress with the Open Data materials and submission status
  - HEPData
  - Docker images for PHENIX (quick overview)
  - Website updates
  - Management of the Analysis Notes

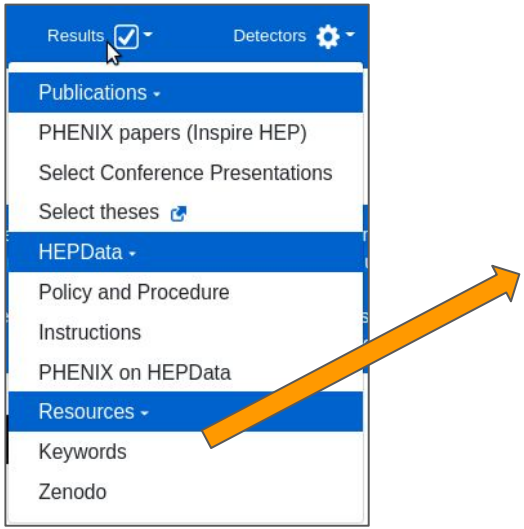
Zenodo (digital repository)

- Using the CERN instance for ~1.5 years
- Uploads happening at a brisk pace
- 310+ items as of today
- Half of them are PhD theses
- Growing number of conference presentations committed to Zenodo
  - Establishing a new workflow for preservation of these materials, instead of the aging legacy web page
- Documents, images, videos, software
- Thanks to Gabor for taking care of conference presentations and to Stacyann Nelson for the theses upload
- More keywords added to the official list as necessary
- Please consider leveraging Zenodo



Keywords

- Maintaining a curated list of keywords is important
- NB. Complex queries involving combination of keywords are possible, and the “elastic search” is built in
- Please check the “Keywords” item under “Results” on the website
- Each keyword is a working link



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 (21 items)

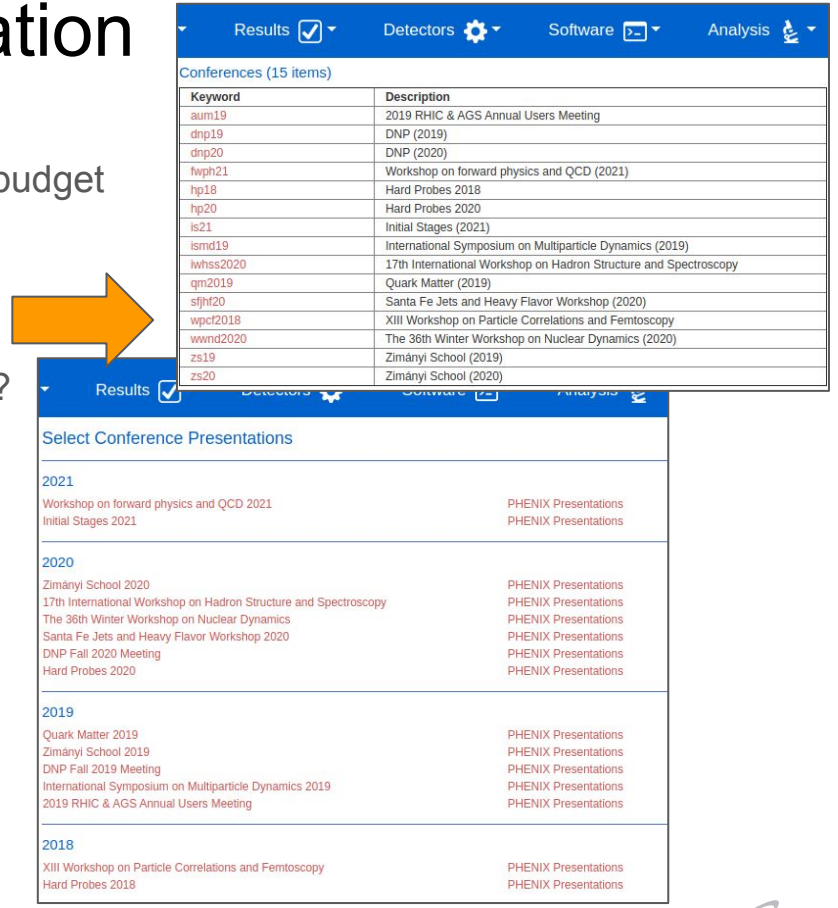
Keyword	Description
aum19	2019 RHIC & AGS Annual Users Meeting
dnp19	DNP (2019)
dnp20	DNP (2020)
fwph21	Workshop on forward physics and QCD (2021)
hp18	Hard Probes 2018
hp20	Hard Probes 2020
icnfp19	International Conference on New Frontiers in Physics 2919
is19	Initial Stages (2019)
is21	Initial Stages (2021)
ismd19	International Symposium on Multiparticle Dynamics (2019)
iwhss2020	17th International Workshop on Hadron Structure and Spectroscopy
qm18	Quark Matter 2018
qm2019	Quark Matter 2019
sfjhf20	Santa Fe Jets and Heavy Flavor Workshop (2020)
sqm16	Strangeness in Quark Matter 2016
sqm17	Strangeness in Quark Matter 2017
sqm19	Strangeness in Quark Matter 2019
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 (93 items)

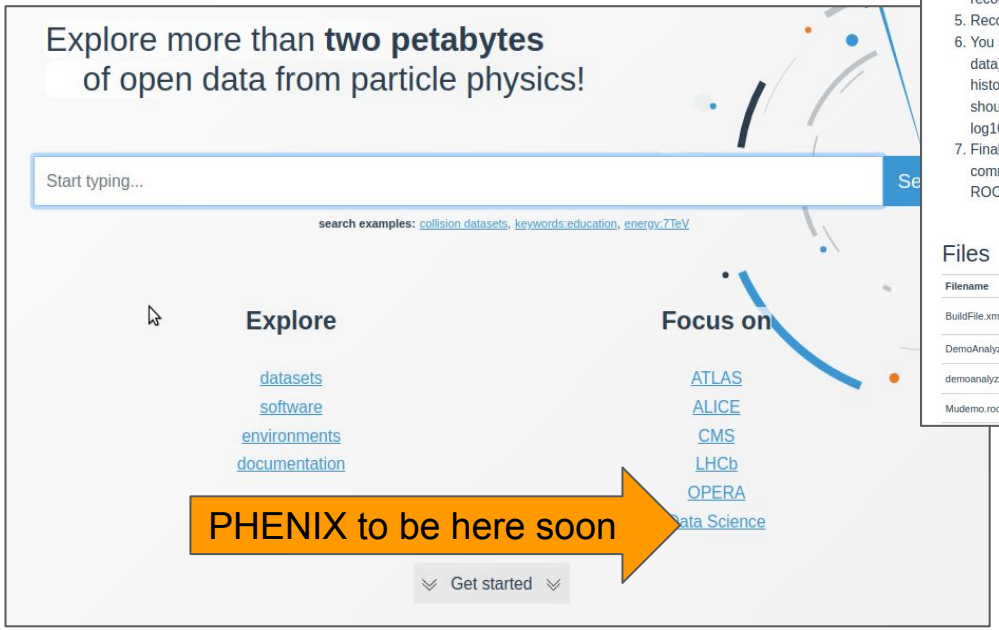
Keyword	Description
3he+au	Helium3-on-gold collisions
anisotropy	Anisotropy
asymmetry	Asymmetry
au+au	Gold-on-gold collisions
azimuthal	Azimuthal
b-meson	B meson

The legacy database situation

- Upload of conference presentation is broken
- Fixing that will likely take too much time for our budget
- For some months now we've been uploading conference materials to Zenodo
- A managed keyword list (including conferences)
- Is this a good time to switch to the new process?



Open Data: a flexible DAP platform



How can you use this?

If you do not have the CERN Virtual Machine for 2010 CMS data installed, follow the instructions in step 1 at How to install a CERN Virtual Machine. Then install and run the Demo (demo analyzer) program following the instructions at How to Test & Validate.

To run the "di-muon spectrum" demo:

1. Create directory `datasets` under `Demo/DemoAnalyzer`.
2. Download the index files for the /Mu/Run2010B-Apr21ReReco-v1/AOD primary datasets and store them in `Demo/DemoAnalyzer/datasets/`.
3. Download the JSON file from CMS Validated Runs and save it to the `Demo/DemoAnalyzer/datasets` directory.
4. Replace the three files `BuildFile.xml`, `demoanalyzer_cfg.py`, `src/DemoAnalyzer.cc` with the ones from this record and read the comments in `DemoAnalyzer.cc` if you want to understand what the program does.
5. Recompile (`scram b`) and rerun (`cmsRun ...`) exactly as shown before in How to Test & Validate.
6. You should get an output file `Mu.root`, which contains histograms for 10000 input events (a small subset of the data). These can be looked at using a ROOT Browser (see above, under Description). The most interesting histogram is `GM_mass_log`. In order to compare it with the invariant-mass spectrum of di-muons in MUO-10-004, it should be viewed with the `logy` option. Of course with 10000 events the comparison is poor, but the J/ψ (at log10(mass)=0.5), Υ (at log10(mass=0.98)) and Z (at log10(mass)=1.95, one event only) peaks should be visible.
7. Finally, to run over more or even the full data, edit the relevant parts of the Python file `demoanalyzer_cfg.py` (see comments therein) and rerun. Add up the output histograms from different (non-overlapping) input index files using ROOT tools.

Files

Filename	Size	
BuildFile.xml	342.0 bytes	Download
DemoAnalyzer.cc	14.3 kB	Download
demoanalyzer_cfg.py	3.4 kB	Download
Mudemo.root	14.6 kB	Download

Open Data - PHENIX engagement

- PHENIX was granted rights to host on the CERN OpenData portal
  - This is a first for an experiment based outside of CERN (and of Europe)
  - The upload mechanism tested, it's simple and reliable (XRootD)
- π^0/γ analysis case captured, materials uploaded - Ntuples, PDF, macros
  - Gabor and Maxim's effort https://github.com/PhenixCollaboration/opendata
- After the initial upload our CERN contacts made a number of suggestions about how to improve the material
- We invested more effort and are close to finalizing this entry
  - Refactored the data
  - Updated the write-up
  - Developed ROOT macros for both legacy (ROOT5) environment and also for ROOT6
  - A Docker image for SL7/ROOT5 has been built

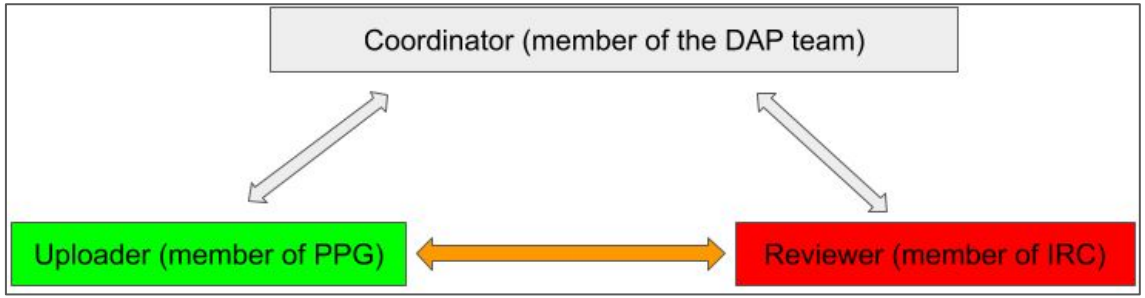
Open Data - going forward

- When time permits, please take a look at the CERN Open Data portal
- A powerful DAP platform which can be leveraged for training and other purposes (future PHENIX Schools?)
- Great publicity for PHENIX
- Question: can we find volunteers to create preserved examples of PHENIX analyses, similar to what we did with π^0/γ?
  - For ROOT-based final stages of analyses creating a package is fairly straightforward
- NB. The LHC experiments are leveraging XRootD/EOS for hosting/shipping data samples - not a requirement but a major convenience. PHENIX does not have that - would be a good idea to get this piece in place going forward

HEPData

- Ongoing work, good participation of the PHENIX community, does take effort
- Contact Maxim and Christine if you have questions or need help

PPG	Contact	Contact e-mail	PPG Uploader name	PPG Uploader e-mail	IRC Reviewer name	IRC Reviewer e-mail	Status/Comments	arXiv	InspireHEP ID	HEPData ID
023	Abdulla Alsayegh	abdullah.alsayegh@protonmail.com					Early draft	308006	625472	
071	K.Smith	kls15k@my.fsu.edu					An older entry, corrections/Conflict: Inspire and Phys.Rev abstracts	801.022	776624	57327
081	Reem Alreshdi	reemalreshidi@outlook.com, cnattra	Zaida Conesa del Valle	zaida.conesa.del.valle@cern.ch			Under sandbox review by Zaida	903.2041	815217	57350
083	C.Nattrass	cnattras@utk.edu	Jiangyong Jia	jiangyong.jia@stonybrook.edu			**Resubmitted and finalized. Done.**		778396	96764
115	C.Nattrass	cnattras@utk.edu					Complete/Done by Dylan Rotunno at UTK			
119	M.Wysocki	matthew.g.wysocki@gmail.com					In preparation	1910.14487	894560	
147	Takahito Todoroki	todoroki@bnl.gov	Takahito Todoroki	todoroki@bnl.gov				1412.1038	1332239	
173	C.Nattrass	cnattras@utk.edu	Takahito Todoroki	todoroki@bnl.gov			Assigned: Christine Nattrass -> Jason Spriggs; decimal places	1803.01749	1658594	
202	Gabor						In preparation, IRC is unclear			
209	S.Zharko	zharkosergey94@gmail.com	S.Zharko	zharkosergey94@gmail.com			Discussion of round off/Pending Sergei's additional edits	1805.04389	1672859	100192 (not public)
210	C.Nattrass	cnattras@utk.edu	C.Nattrass	cnattras@utk.edu	A.Hodges	ahodges21@student.gsu.edu	**PNG's corrected for quality. Done.**	2005.14270	1798493	101752
235	N.Lewis	nialewis@umich.edu	Nicole Lewis	nialewis@umich.edu	Xiaochun He	xhe@gsu.edu	Comment period ended			



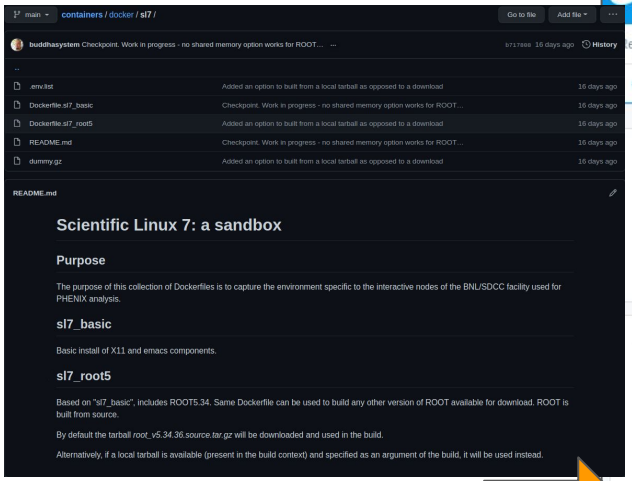
Docker: the purpose and the challenges

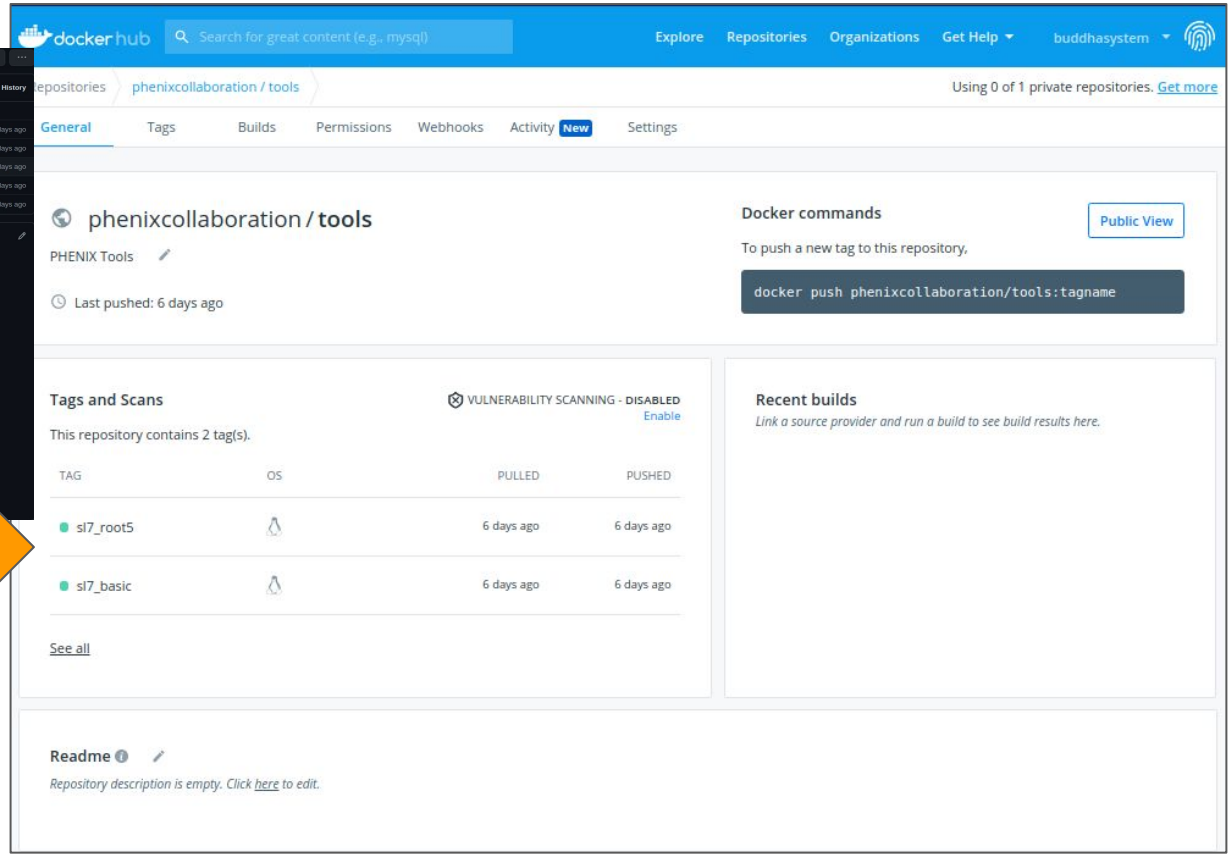
- Keep/capture snapshots of individual stages of analyses - with proper containers a lot can be done on laptops/workstations - preserve the software environment
  - cf. legacy ROOT, compilers as they exist on interactive and batch nodes now
- Aid in capturing workflows with REANA
- Challenge: building the software - regardless whether it's for Docker images or other purposes - if you can build it, you can containerize it
  - Build procedures in PHENIX became rather monolithic and complex over time
  - Looking into it
- Copying over .so's is an option if you can be sure of all dependencies
- CVMFS might be an option as well

Docker

- Created a new GitHub repository for PHENIX: “containers”
  - e.g. for Dockerfiles
  - For management of materials needed to build various images - Docker, Singularity as needed
- Current ROOT 6+ versions are provided by the ROOT team - but not ROOT5
  - https://hub.docker.com/u/rootproject - runs out of the box
- Built a ROOT 5.34/36 image - the PHENIX standard
  - Configuration kept on GitHub
  - PHENIX organization created on Docker Hub, custom images built and pushed to the repo
  - They also include compatible versions of gcc and g++, same as on interactive nodes
  - Available to anyone
- Documented added as the new pages on the website

Docker - GitHub and Docker Hub

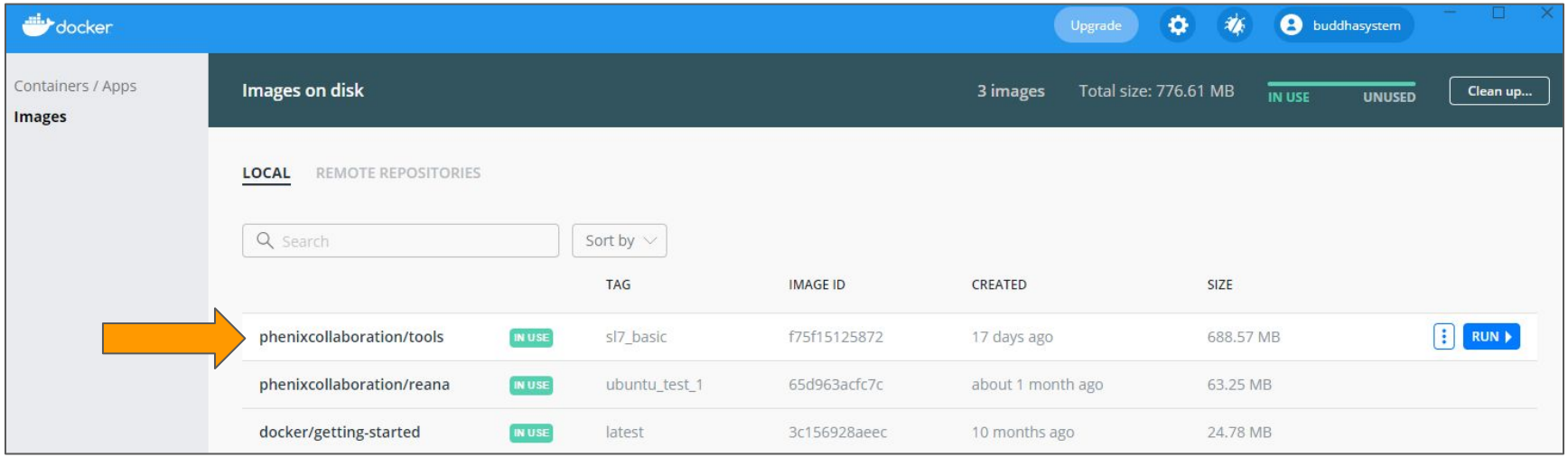




Can run in the familiar SL7 on any version of Linux, MacOS and Windows

Docker on Windows

- Reasonable installation procedure (a couple of updates and restarts)
- Runs under WSL 2 - a Linux environment - closer to the kernel than a VM
- Features a desktop application plus full command-line functionality

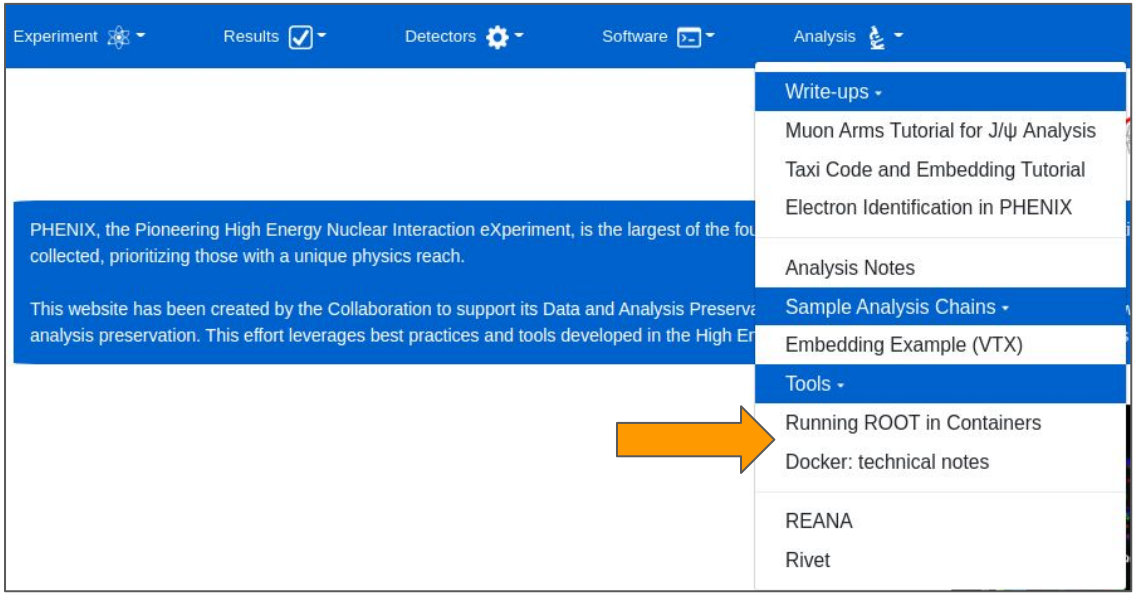


Docker on Mac

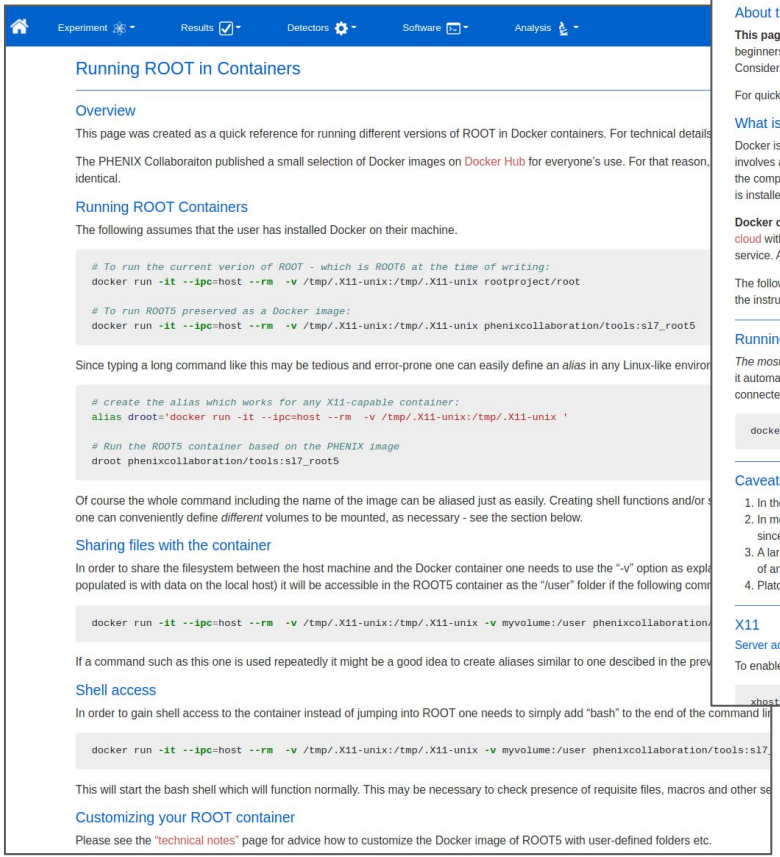
- I don’t have a Mac so can’t report on experience
- Many users have Macs - any volunteers to try it out?
- Can have a complete SL7 environment without installing complex software
  - Same gcc, g++, emacs, xterm, ROOT as on the RACF nodes

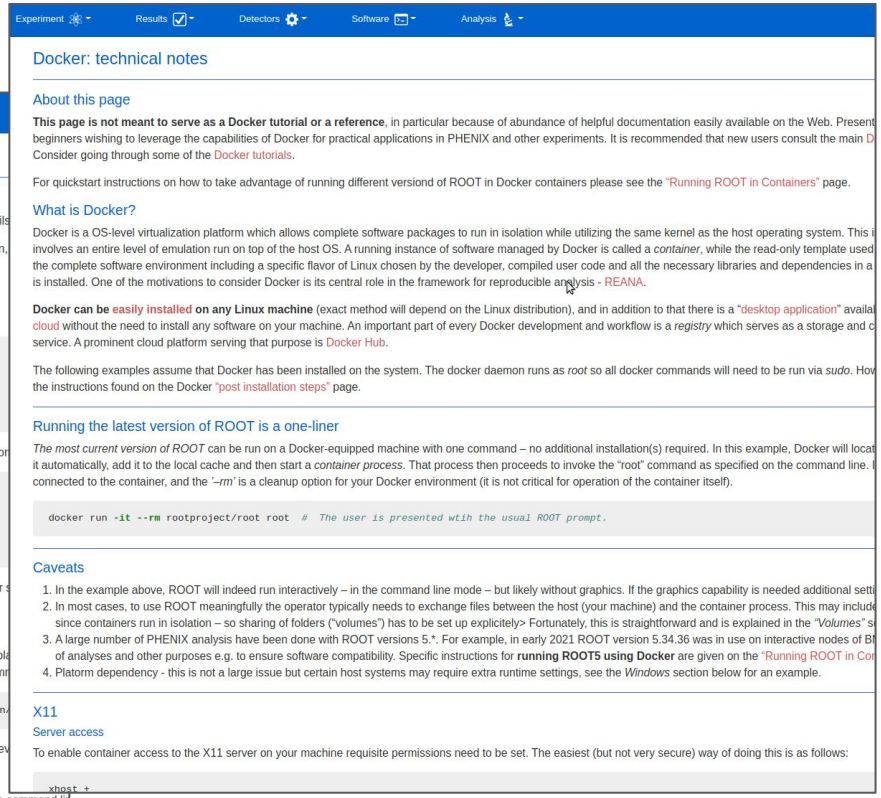
Website updates

- Docker pages created, fair amount of material added
  - Quick start + Technical details



Website: Docker Pages





Analysis notes

- Long-standing requirements for data privacy and access controls
- Discoverability for authorized users, cross-reference
- Aging infrastructure (legacy web pages) - need durable long-term solutions

Analysis notes: all options (details on the next slide)

- Any type of file sharing option - with encryption
  - Passwords can be circulated to select participants only
  - Finding a truly portable solution may be a bit of a challenge, openssl is a strong contender (all platforms)
- Zenodo - private access option
  - Access on demand, decided by the PHENIX Zenodo curators
  - The only solution offering proper built-in indexing and search capabilities
- GitHub - a private repository
  - Accessible to users on a managed list
  - GitHub tags can be used for indexing (like keywords)
- BNLbox
  - Broadly speaking, an equivalent of Dropbox with vastly larger storage available
  - Web UI
  - File upload and download using a CLI script is possible
  - A fairly capable access control system

Analysis notes options: the shortlist

- Zenodo - private access option
  - Access on demand, decided by the PHENIX Zenodo curators
  - Per-person, per item action - this can only work if there are <20 interested people or so for a given note
  - Designed for “close to final” or “final” versions of documents i.e. not as a workflow mechanism
  - The only solution offering proper built-in indexing and search capabilities
  - References to prior notes on same topics can be handled organically
  - No convenient folder structure i.e. directory trees
- GitHub - a private repository
  - PHENIX does have an organization on GitHub i.e. everything can be properly managed
  - Accessible to users on a managed list
  - Easier management of access, no individual action per item/person
  - Flexible folder structure
  - Designed for revisions/workflows
  - GitHub tags can be used for indexing (like keywords)