Data & Analysis Preservation: current work items

Maxim Potekhin
Nuclear and Particle Physics Software Group

BROOKHAVEN NATIONAL LABORATORY

PHENIX

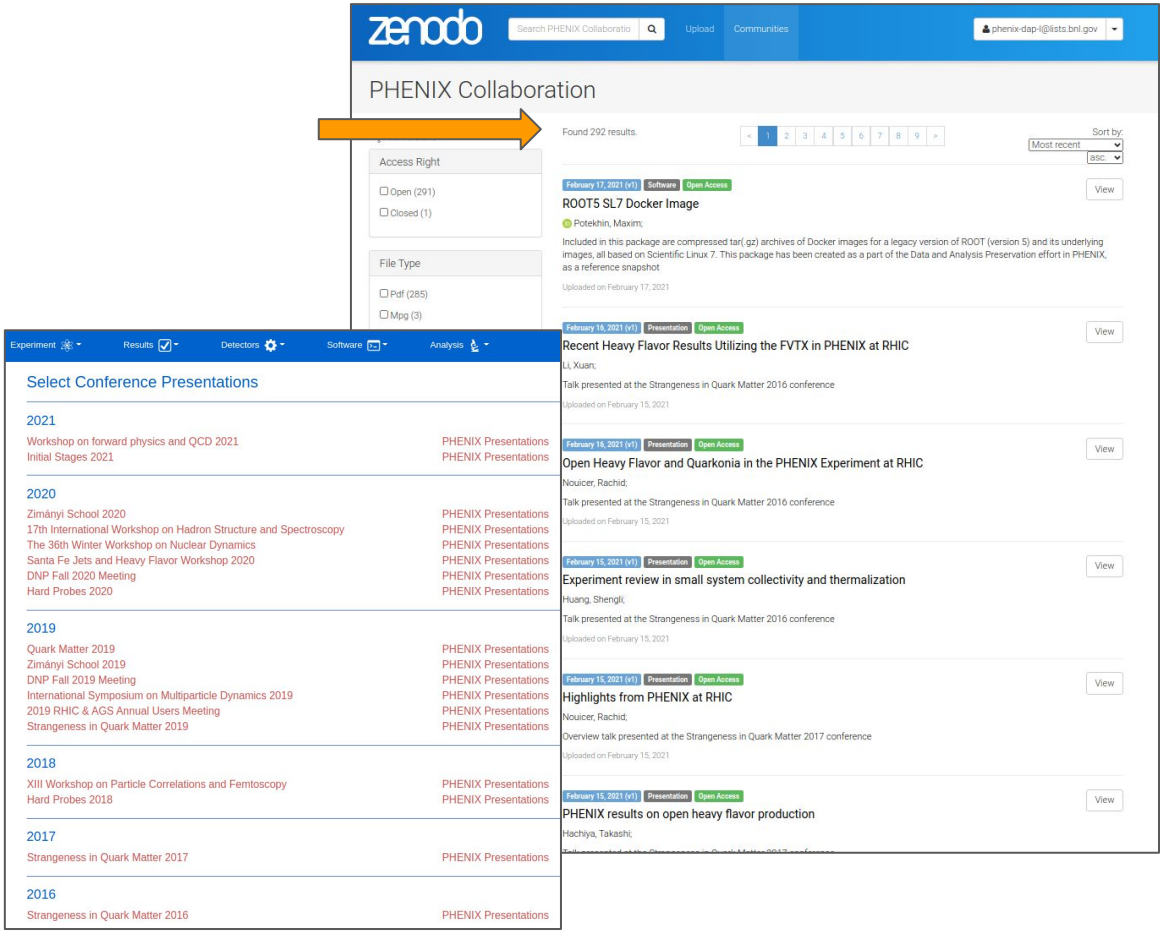
PHENIX DAP Meeting
02/25/2021

Overview

- Zenodo
- The legacy DB situation
- GitHub repo plans
- OpenData
- HEPData
- Docker development
- Website updates
- Analysis notes redux

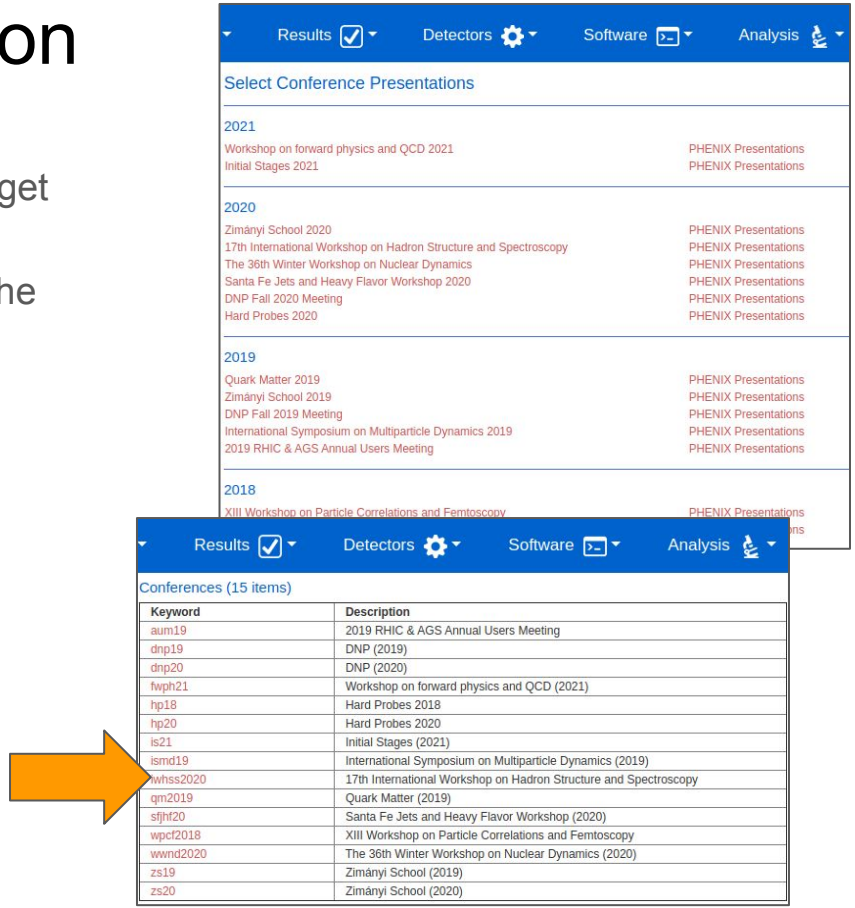
Zenodo

- Uploads happening at a brisk pace
- 300+ items as of today
- Thanks to Gabor for taking care of conference presentations and to Stacyann for the theses
- More keywords added to the official list as necessary



The legacy database situation

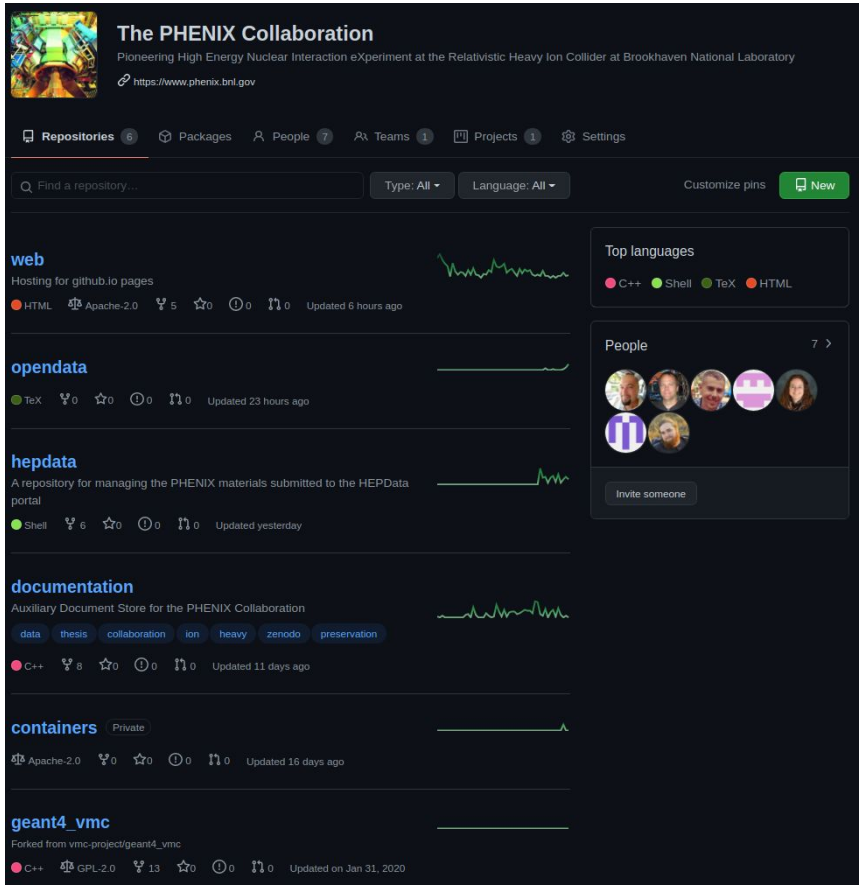
- Upload of conference presentation is broken
- Fixing that will likely take too much time for our budget
- For some months now we've been uploading conference materials to Zenodo while maintaining the keyword list (including keywords for conferences)
- A lot of effort invested by Gabor
- Is this a good time to switch to the new process?
- It's not difficult to set up a workflow with GitHub, BNLbox for revisions etc
  - ...since uploading drafts to Zenodo is not a good idea



GitHub: the “documentation” repo

- The repo was created a long time ago as a general container for all kinds of documents used in the PHENIX DAP effort
- The HEPData component was factored out to a separate repo to make it lightweight and agile - this is a team effort and multiple people are doing the “git pull”
- Now, upload and backup of the PHENIX theses is close to finish
  - Many thanks to Stacyann Nelson
- The resulting volume of the theses folder in the “documentation” repository is somewhat larger than expected - O(GB) - and makes it harder to access smaller sets of other types of files if needed
- It makes sense to create a dedicated repo for the theses, stay tuned
  - Will make decision/proceed after all theses are uploaded

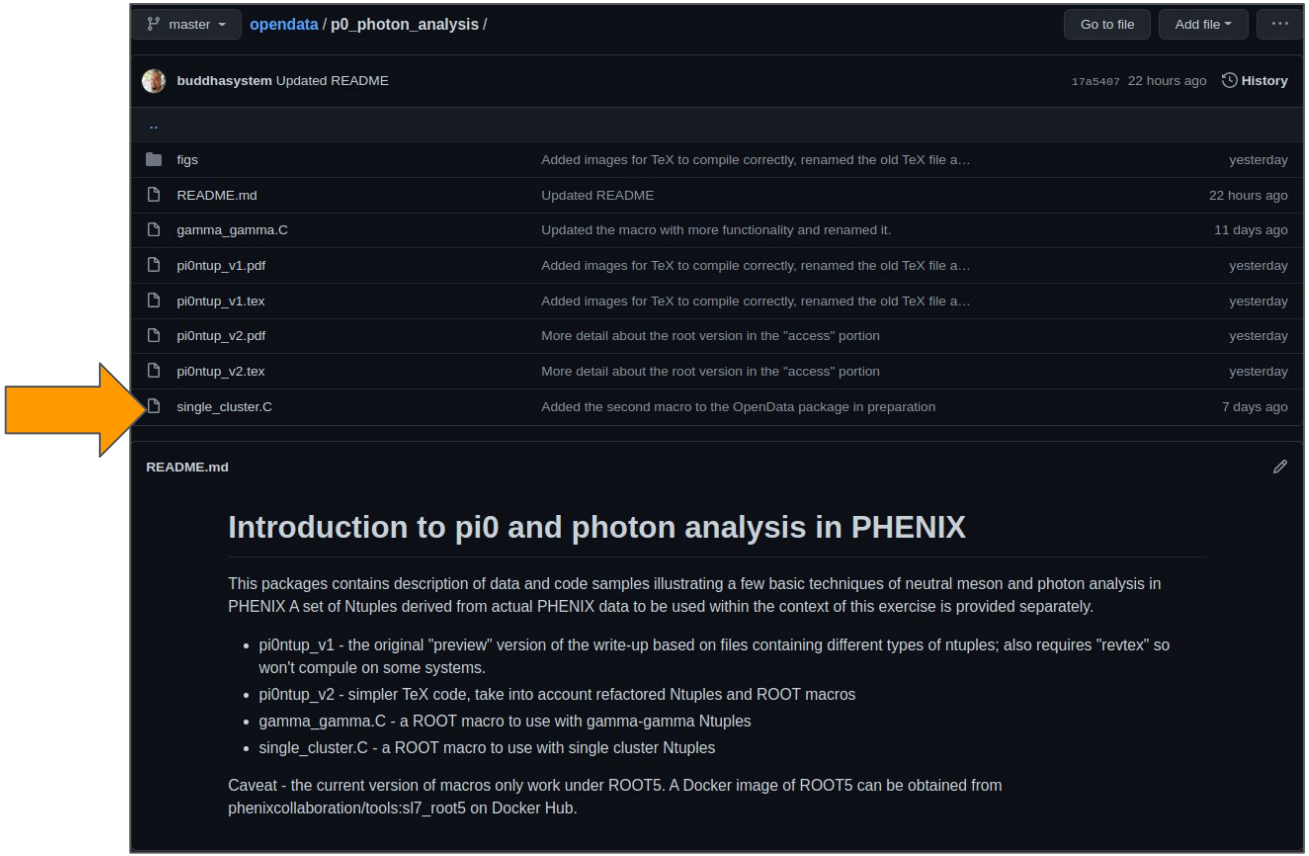
PHENIX on GitHub



Open Data - the “π^0 and ɣ analysis” entry

- Recall: uploaded to Open Data in the of Fall’2020, comments received
  - The upload mechanism and auth/auth tested and works (XRootD)
- Survey of datasets from other experiments (i.e. the “dataset semantics” feature)
  - A built-in feature of the OpenData pages to document datasets, one table per page
- Heterogeneous datasets have to be documented in the textual description
  - Gabor’s original files had two different ntuples per file
  - Site admins suggested refactoring data (not mandatory but desirable)
- Thanks to Gabor making that extra step of splitting data
- Another recommendation from OpenData - have standalone code as opposed to snippets in the PDF
  - Has now been done, ROOT macros created and tested
- This required changes in the document - now also completed

Our OpenData entry on GitHub



master ▾ opendata / p0_photon_analysis / Go to file Add file ▾ ···

buddhasystem Updated README 17a5407 22 hours ago History

..

figs	Added images for TeX to compile correctly, renamed the old TeX file a…	yesterday
README.md	Updated README	22 hours ago
gamma_gamma.C	Updated the macro with more functionality and renamed it.	11 days ago
pi0ntup_v1.pdf	Added images for TeX to compile correctly, renamed the old TeX file a…	yesterday
pi0ntup_v1.tex	Added images for TeX to compile correctly, renamed the old TeX file a…	yesterday
pi0ntup_v2.pdf	More detail about the root version in the "access" portion	yesterday
pi0ntup_v2.tex	More detail about the root version in the "access" portion	yesterday
single_cluster.C	Added the second macro to the OpenData package in preparation	7 days ago

README.md

Introduction to pi0 and photon analysis in PHENIX

This packages contains description of data and code samples illustrating a few basic techniques of neutral meson and photon analysis in PHENIX A set of Ntuples derived from actual PHENIX data to be used within the context of this exercise is provided separately.

- pi0ntup_v1 - the original "preview" version of the write-up based on files containing different types of ntuples; also requires "revtex" so won't compule on some systems.
- pi0ntup_v2 - simpler TeX code, take into account refactored Ntuples and ROOT macros
- gamma_gamma.C - a ROOT macro to use with gamma-gamma Ntuples
- single_cluster.C - a ROOT macro to use with single cluster Ntuples

Caveat - the current version of macros only work under ROOT5. A Docker image of ROOT5 can be obtained from phenixcollaboration/tools:sl7_root5 on Docker Hub.

OpenData: next steps

- Request to Gabor - is it possible to create macros compatible with ROOT6?
- Final overview of the new version of the write-up
- Do we have any other analysis cases suitable for Open Data submission?
- Combination of data hosting/access plus code/documentation makes it very suitable for training and educational purposes cf. PHENIX Schools, while also being a part of DAP
- Can we get any volunteers to consider making more packages?
- Caveat: to be useable the software included in a package must be portable, which is - either amenable to a build, or to containerization; cf. in our first entry we used ROOT macros which are portable by definition

HEPData

- Ongoing HEPData preparation and management
  - Quite a few items are being worked on, a sizable work item, takes time
  - Our new workflow/review process does work:
    - Uploader <--> Reviewer
    - GitHub at the center
  - Please see the spreadsheet (reformatted a little bit for compactness) - also next slide: https://docs.google.com/spreadsheets/d/1rABxzuM-h9Rukz08ut_m8xnMo0B_J1LKre8bM7B7264/edit?usp=sharing
- Progress in the past three weeks:
  - PPG023, PPG235 - initial stubs/sandboxing
  - PPG083 - correction of an existing prior HEPData entry - done
  - PPG081 - correction of an existing prior HEPData entry - work in progress
  - Scientific notation/decimal places in PPG209 - work in progress
  - PPG210 - done, made corrections to PNG images

HEPData: the spreadsheet

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			

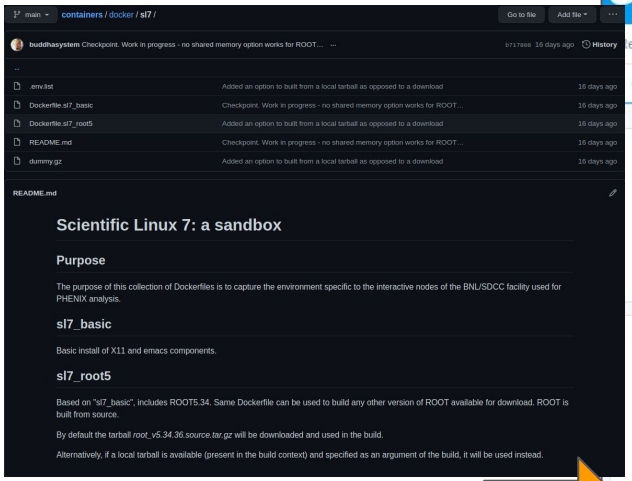
Docker: the purpose and the challenges

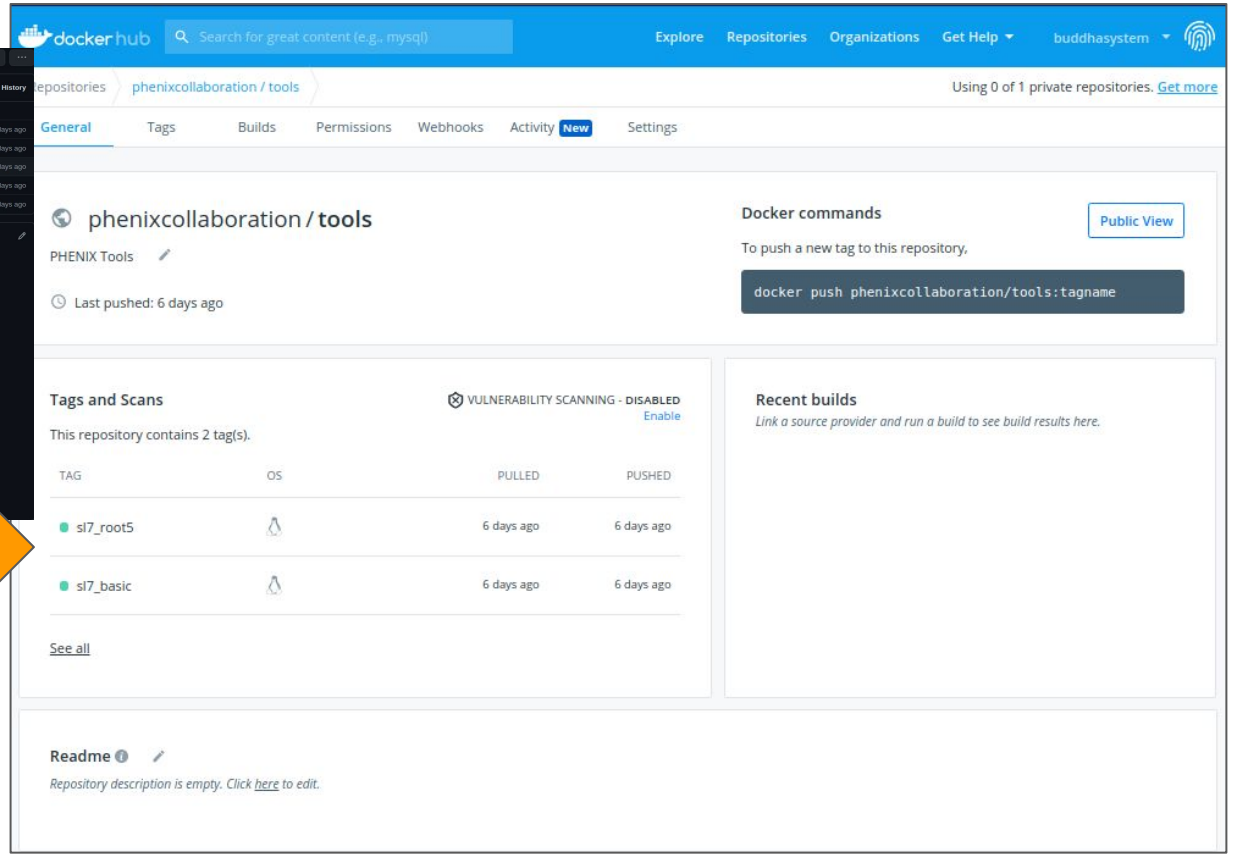
- To keep/capture snapshots of individual stages of analyses
- To aid in capturing workflows with REANA
- To preserve the software environment (at least some of it)
- Challenge: building the software - regardless whether it's for Docker images or other purposes
- If you can build it, you can containerize it

Docker

- Created a new GitHub repository for PHENIX: “containers”
  - For management of Dockerfiles and configuration materials needed to build various images
  - Keeping it as “private” for now - let me know if you need access
- Current ROOT 6+ versions are provided by the ROOT team - but not ROOT5
  - https://hub.docker.com/u/rootproject
- Built a ROOT 5.34/36 image - the current standard on interactive nodes
  - Configuration kept on GitHub
  - PHENIX organization created on Docker Hub
  - Custom images built and deposited to the PHENIX repository on Docker Hub
  - Also uploaded to Zenodo (as tarballs)
  - They also include compatible versions of gcc and g++, same as on interactive nodes
  - Available to anyone
- A few caveats (X11 etc) documented in detail on the new pages on the website
  - Please see slides on the topic of the website, later in this presentation

Docker - GitHub and Docker Hub

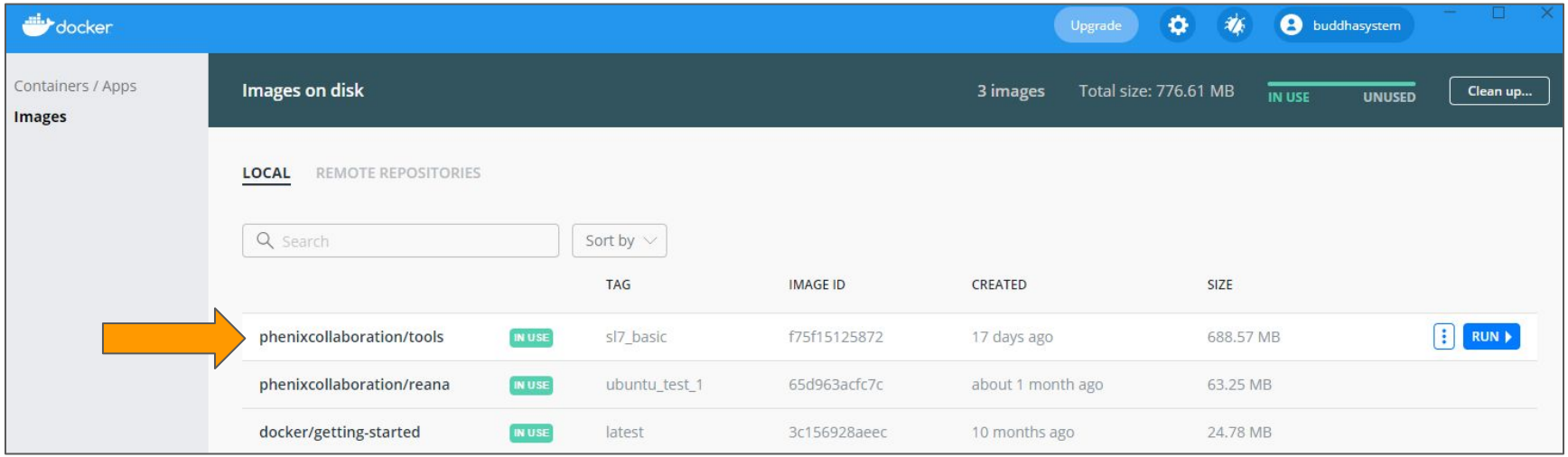




Can run in the familiar SL7 on any version of Linux, MacOS and Windows

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

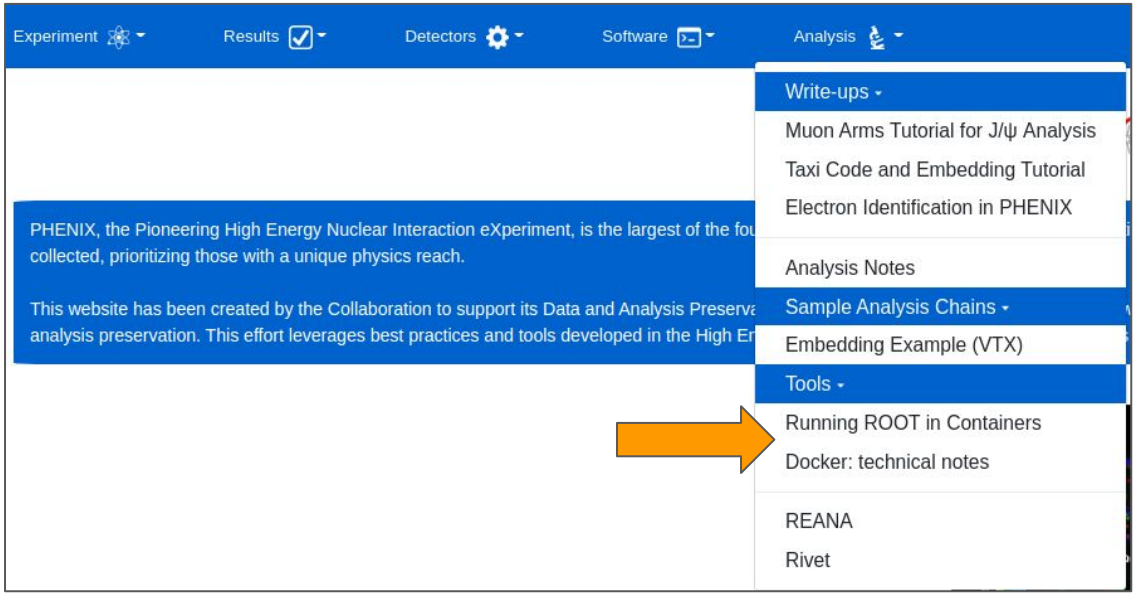
- I don’t have a Mac so can’t report on experience
- Many users have Macs
- Volunteers to try it out?
- Can have a complete SL7 environment without installing complex software
  - Same gcc, g++, emacs, xterm, ROOT as on the RACF nodes

Docker on Windows - CLI, ROOT

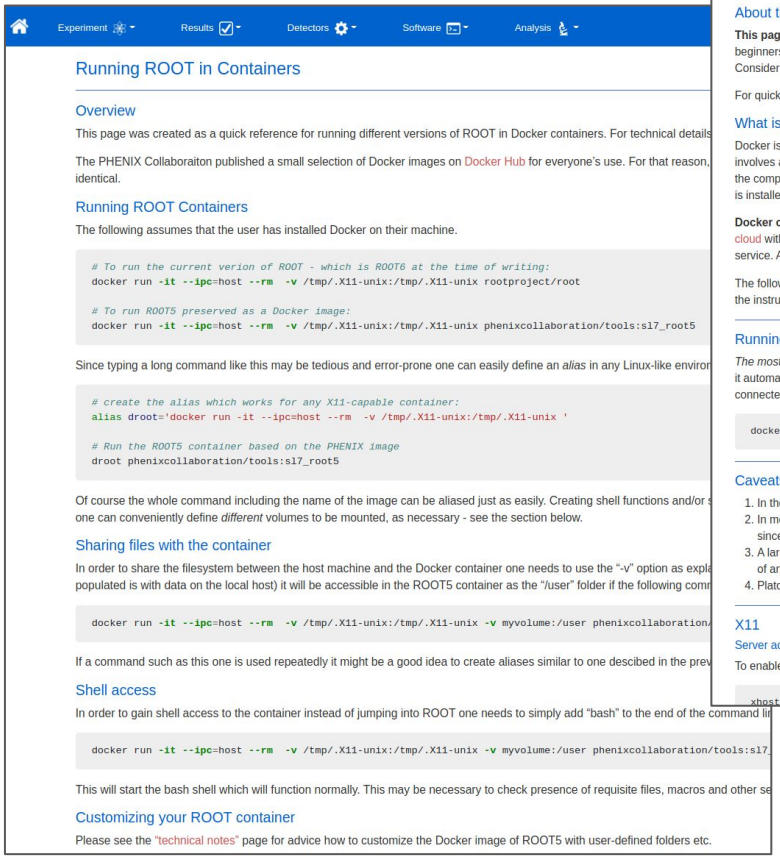
```
eabb785359ea: Pull complete
Digest: sha256:368dc0b59de676cfffacb468cb9eaed54fdc54f700842a71b11d0c5bdaac7a6a
Status: Downloaded newer image for phenixcollaboration/tools:sl7_root5
docker.io/phenixcollaboration/tools:sl7_root5

 24/02/2021  19:52.19  /home/mobaxterm  docker image ls
REPOSITORY                  TAG             IMAGE ID       CREATED         SIZE
phenixcollaboration/tools   sl7_basic       f75f15125872   2 weeks ago     689MB
phenixcollaboration/tools   sl7_root5       156b05cbb23c   2 weeks ago     3.16GB
phenixcollaboration/reana   ubuntu_test_1   65d963acfc7c   6 weeks ago     63.3MB
docker/getting-started      latest          3c156928aeec   10 months ago   24.8MB

 24/02/2021  19:52.55  /home/mobaxterm  docker run -it --rm phenixcollaboration/tools:root5
Unable to find image 'phenixcollaboration/tools:root5' locally
docker: Error response from daemon: manifest for phenixcollaboration/tools:root5 not found: manifest unknown: manifest
unknown.
See 'docker run --help'.

 24/02/2021  19:54.23  /home/mobaxterm  docker run -it --rm phenixcollaboration/tools:sl7_root5
  ****************************************
  *                                      *
  *       W E L C O M E  to  R O O T      *
  *                                      *
  *   Version   5.34/36      5 April 2016  *
  *                                      *
  *  You are welcome to visit our Web site  *
  *         http://root.cern.ch          *
  *                                      *
  ****************************************

ROOT 5.34/36 (v5-34-36@v5-34-36, Apr 05 2016, 10:25:45 on linuxx8664gcc)

CINT/ROOT C/C++ Interpreter version 5.18.00, July 2, 2010
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.
root [0] tb=TBrowser();
root [1]
```

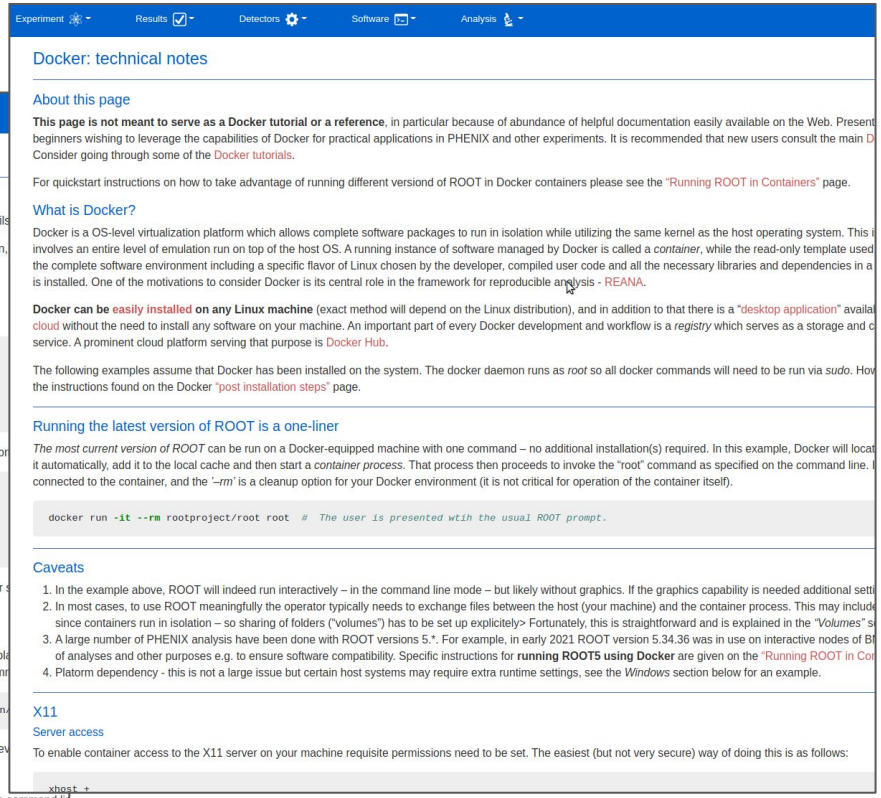
Website updates

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



Website: Docker Pages





Analysis notes options: a redux

- 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

Status and Plans

- Look at the overlap of the PHENIX School and DAP work
- Open Data - make our first entry final
- Ongoing HEPData work
- The hard part - build useable Docker images for DAP - engagement with the Collaboration is required
  - Prerequisite to REANA