

Data & Analysis Preservation: current work items

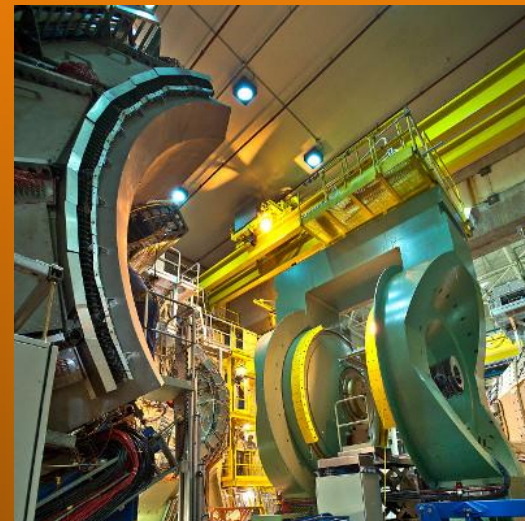
Maxim Potekhin

Nuclear and Particle Physics Software Group



PHENIX DAP Meeting

02/25/2025



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

Select Conference Presentations

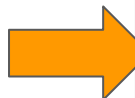
Year	Conference/Workshop	Link
2021	Workshop on forward physics and QCD 2021 Initial Stages 2021	PHENIX Presentations PHENIX Presentations
2020	Zimányi School 2020 17th International Workshop on Hadron Structure and Spectroscopy The 36th Winter Workshop on Nuclear Dynamics Santa Fe Jets and Heavy Flavor Workshop 2020 DNP Fall 2020 Meeting Hard Probes 2020	PHENIX Presentations PHENIX Presentations PHENIX Presentations PHENIX Presentations PHENIX Presentations PHENIX Presentations
2019	Quark Matter 2019 Zimányi School 2019 DNP Fall 2019 Meeting International Symposium on Multiparticle Dynamics 2019 2019 RHIC & AGS Annual Users Meeting Strangeness in Quark Matter 2019	PHENIX Presentations PHENIX Presentations PHENIX Presentations PHENIX Presentations PHENIX Presentations PHENIX Presentations
2018	XIII Workshop on Particle Correlations and Femtoscopy Hard Probes 2018	PHENIX Presentations PHENIX Presentations
2017	Strangeness in Quark Matter 2017	PHENIX Presentations
2016	Strangeness in Quark Matter 2016	PHENIX Presentations

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

Results <input checked="" type="checkbox"/> Detectors Software Analysis		
Select Conference Presentations		
2021		
Workshop on forward physics and QCD 2021		PHENIX Presentations
Initial Stages 2021		PHENIX Presentations
2020		
Zimányi School 2020		PHENIX Presentations
17th International Workshop on Hadron Structure and Spectroscopy		PHENIX Presentations
The 36th Winter Workshop on Nuclear Dynamics		PHENIX Presentations
Santa Fe Jets and Heavy Flavor Workshop 2020		PHENIX Presentations
DNP Fall 2020 Meeting		PHENIX Presentations
Hard Probes 2020		PHENIX Presentations
2019		
Quark Matter 2019		PHENIX Presentations
Zimányi School 2019		PHENIX Presentations
DNP Fall 2019 Meeting		PHENIX Presentations
International Symposium on Multiparticle Dynamics 2019		PHENIX Presentations
2019 RHIC & AGS Annual Users Meeting		PHENIX Presentations
2018		
XIII Workshop on Particle Correlations and Femtoscopy		PHENIX Presentations

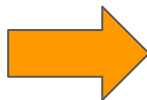
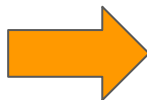
Results <input checked="" type="checkbox"/> Detectors Software Analysis		
Conferences (15 items)		
Keyword	Description	
aum19	2019 RHIC & AGS Annual Users Meeting	
dnp19	DNP (2019)	
dnp20	DNP (2020)	
fwph21	Workshop on forward physics and QCD (2021)	
hp18	Hard Probes 2018	
hp20	Hard Probes 2020	
is21	Initial Stages (2021)	
ismd19	International Symposium on Multiparticle Dynamics (2019)	
whss2020	17th International Workshop on Hadron Structure and Spectroscopy	
qm2019	Quark Matter (2019)	
sfjhf20	Santa Fe Jets and Heavy Flavor Workshop (2020)	
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)	



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

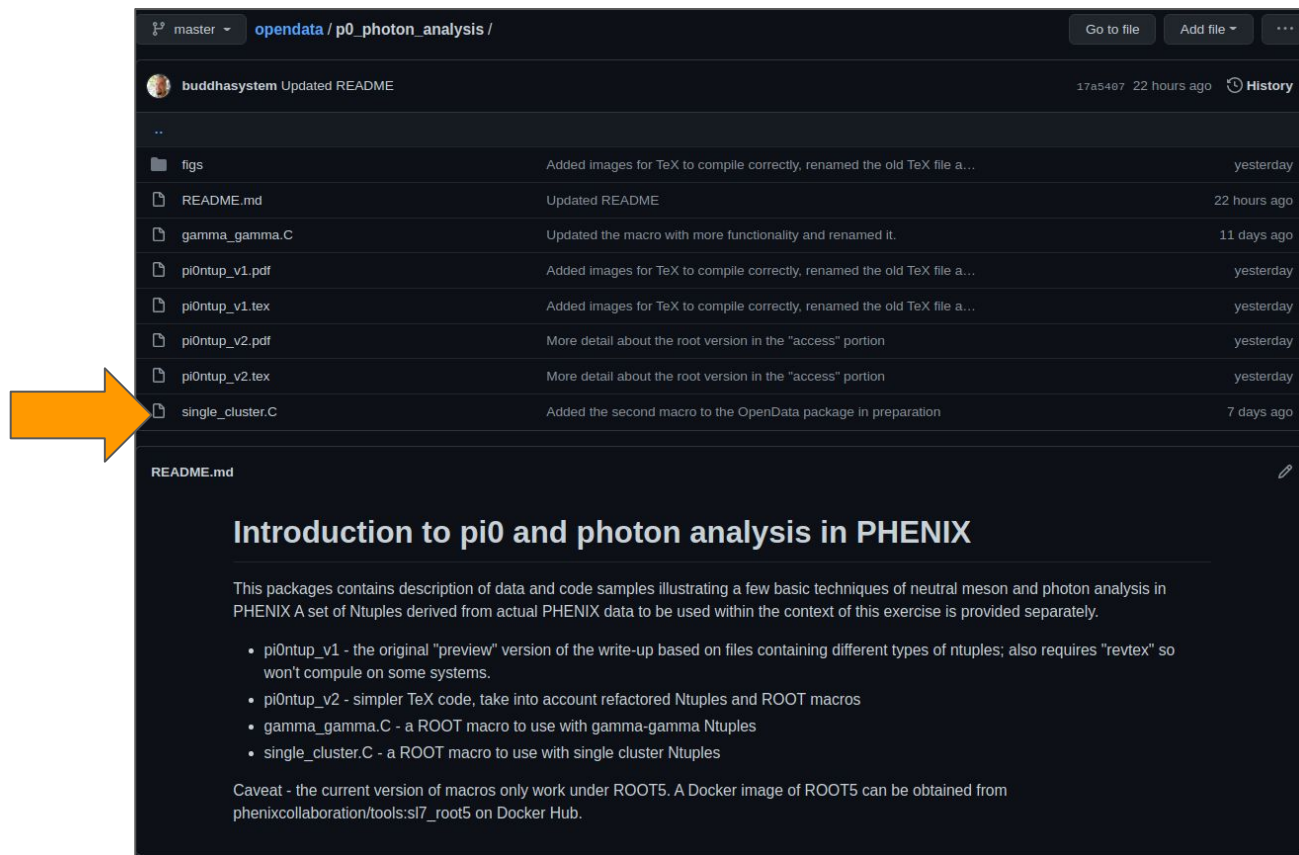


The screenshot shows the GitHub profile for 'The PHENIX Collaboration'. The profile header includes the organization's name, a description: 'Pioneering High Energy Nuclear Interaction eXperiment at the Relativistic Heavy Ion Collider at Brookhaven National Laboratory', and a website link: 'https://www.phenix.bnl.gov'. Below the header, there are navigation tabs for Repositories (6), Packages, People (7), Teams (1), Projects (1), and Settings. A search bar and filters for repository type and language are present. The main content area lists several repositories: 'web' (HTML, Apache-2.0, 5 forks, 0 stars, updated 6 hours ago), 'opendata' (TeX, 0 forks, 0 stars, updated 23 hours ago), 'hepdata' (Shell, 6 forks, 0 stars, updated yesterday), 'documentation' (C++, 8 forks, 0 stars, updated 11 days ago), 'containers' (Private, Apache-2.0, 0 forks, 0 stars, updated 16 days ago), and 'geant4_vmc' (C++, GPL-2.0, 13 forks, 0 stars, updated on Jan 31, 2020). On the right side, there are sections for 'Top languages' (C++, Shell, TeX, HTML) and 'People' (7 members, with an 'Invite someone' button).

Open Data - the “ π^0 and γ 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



The screenshot shows a GitHub repository interface. At the top, the repository name is 'opendata / p0_photon_analysis /'. Below this, a commit by 'buddhasystem' is shown, dated '17a5467 22 hours ago'. A list of files is displayed, including 'figs', 'README.md', 'gamma_gamma.C', 'pi0ntup_v1.pdf', 'pi0ntup_v1.tex', 'pi0ntup_v2.pdf', 'pi0ntup_v2.tex', and 'single_cluster.C'. An orange arrow points to the 'README.md' file. Below the file list, the content of 'README.md' is shown, starting with the title 'Introduction to pi0 and photon analysis in PHENIX'.

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 compile 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](https://github.com/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	kis15k@my.fsu.edu					An older entry, corrections/Conflict: Inspire and Phys.Rev abstracts	801.022	776624	57327
081	Reem Alreshdi	reemalreshdi@outlook.com, cnattras	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@bni.gov	Takahito Todoroki	todoroki@bni.gov				1412.1038	1332239	
173	C.Nattrass	cnattras@utk.edu	Takahito Todoroki	todoroki@bni.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

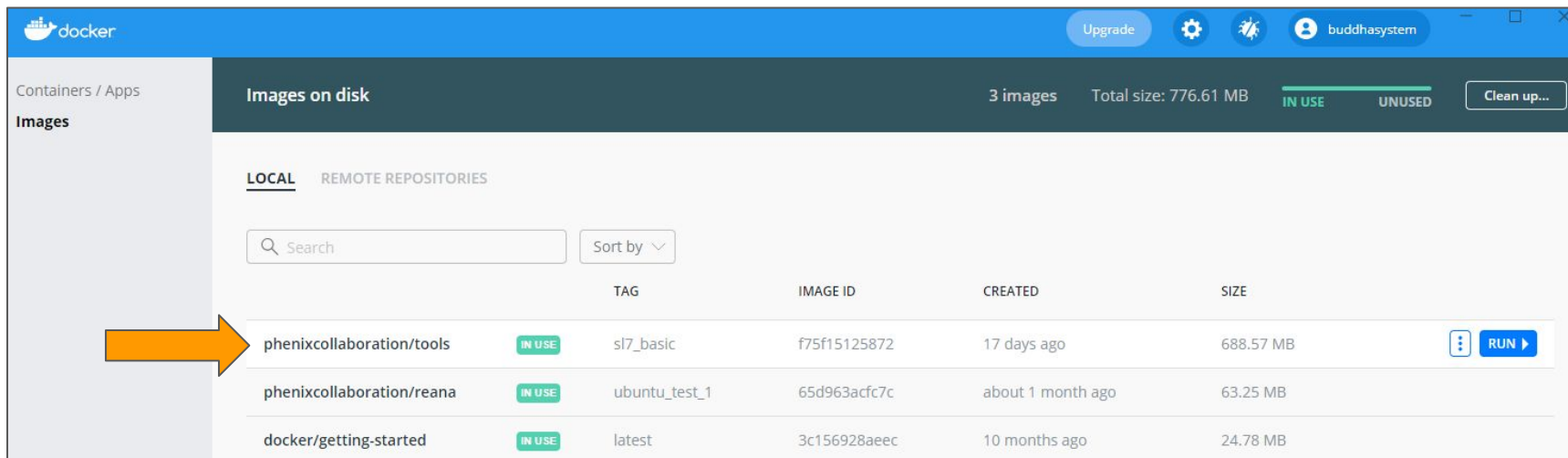
The image shows a Docker Hub repository page for 'phenixcollaboration / tools' and a terminal window displaying the contents of a Dockerfile. The terminal window is titled 'containers / docker / sl7 /' and shows a list of Dockerfiles: 'ami-list', 'Dockefile.sl7_basic', 'Dockefile.sl7_root5', 'README.md', and 'dummy.gz'. The 'README.md' file is selected, showing its content: 'Scientific Linux 7: a sandbox', 'Purpose', 'sl7_basic', and 'sl7_root5'. The 'sl7_root5' section mentions 'Based on "sl7_basic", includes ROOT5.34. Same Dockerfile can be used to build any other version of ROOT available for download. ROOT is built from source.' and 'By default the tarball root_v5.34.36.source.tar.gz will be downloaded and used in the build. Alternatively, if a local tarball is available (present in the build context) and specified as an argument of the build, it will be used instead.' An orange arrow points from the 'sl7_root5' section of the README to the 'sl7_root5' tag in the Docker Hub repository page. The Docker Hub page shows the repository name 'phenixcollaboration / tools', the tag 'PHENIX Tools', and the last pushed time '6 days ago'. It also displays a table of tags and scans, showing two tags: 'sl7_root5' and 'sl7_basic', both pushed 6 days ago. The 'Recent builds' section is empty, and the 'Readme' section is also empty.

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

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



The screenshot shows the Docker Desktop interface. The top bar is blue with the Docker logo, an 'Upgrade' button, and user information for 'buddhasystem'. The main area is titled 'Images on disk' and shows 3 images with a total size of 776.61 MB. A progress bar indicates 'IN USE' and 'UNUSED' status. Below this, there are tabs for 'LOCAL' and 'REMOTE REPOSITORIES'. A search bar and a 'Sort by' dropdown are present. The table lists three images:


		TAG	IMAGE ID	CREATED	SIZE	
phenixcollaboration/tools	IN USE	sl7_basic	f75f15125872	17 days ago	688.57 MB	[Menu] RUN ▶
phenixcollaboration/reana	IN USE	ubuntu_test_1	65d963acfc7c	about 1 month ago	63.25 MB	
docker/getting-started	IN USE	latest	3c156928aeec	10 months ago	24.78 MB	

An orange arrow points to the 'phenixcollaboration/tools' image.

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: sha256:368dc0b59de676cffffac6468cb9eae54fdc54f700842a71b11d0c5bdaac7a6a
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'.

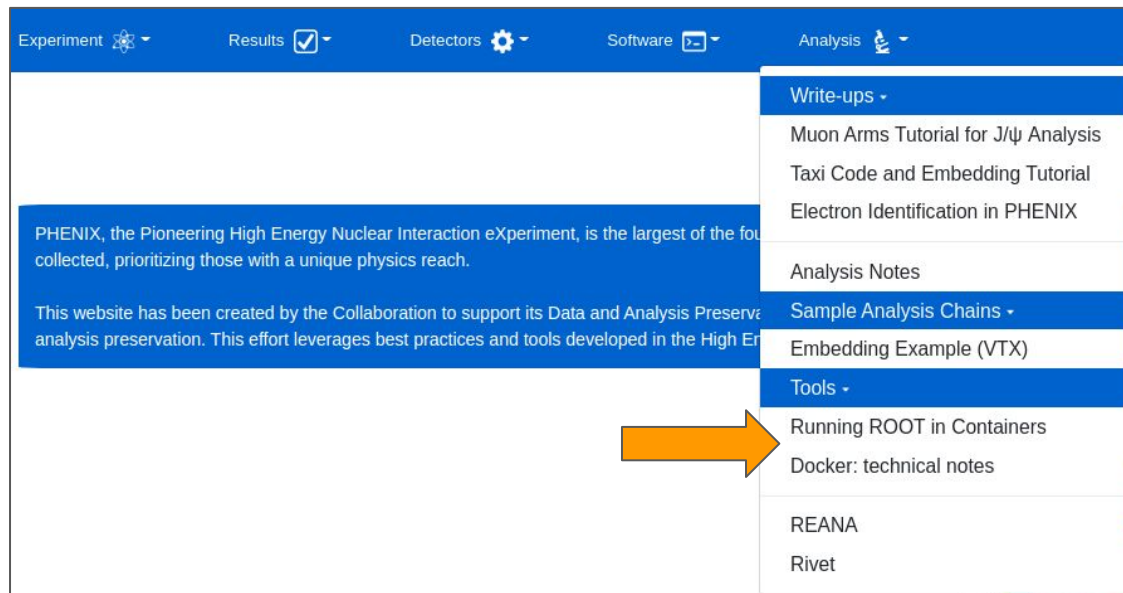
24/02/2021 19:54:23 /home/mobaxterm docker run -it --rm phenixcollaboration/tools:sl7_root5
*****
*
*      W E L C O M E  t o  R O O T      *
*
*   Version   5.34/36      5 April 2016   *
*
*   You are welcome to visit our Web site *
*   http://root.cern.ch *
*
*****

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

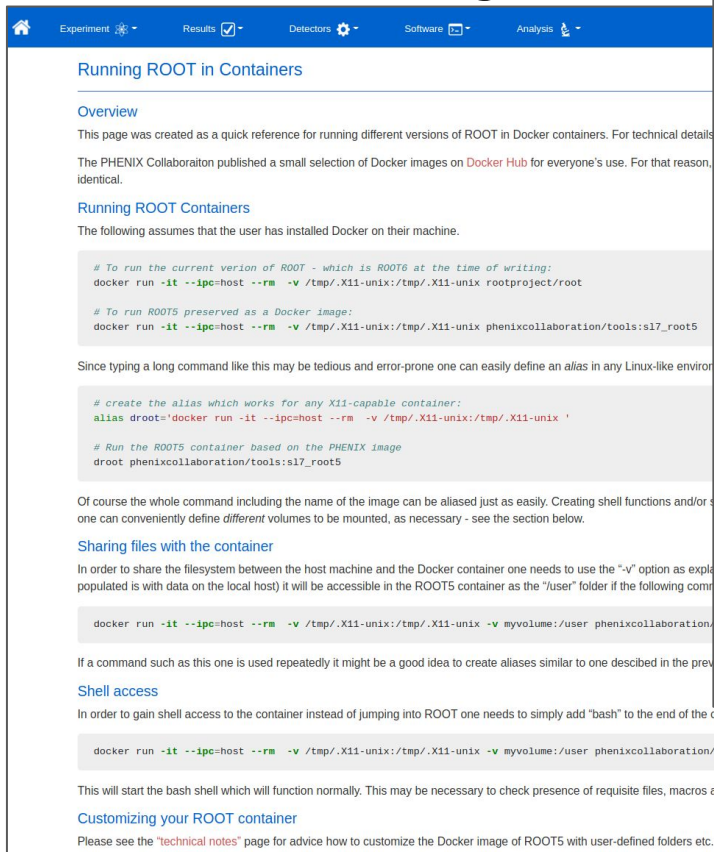
CINT/ROOT C/C++ Interpreter version 5.18.00, July 2, 2010
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.
root [0] tb=TBrowser();
root [1] █
```

Website updates

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



Website: Docker Pages



The screenshot shows the 'Running ROOT in Containers' page. It includes an overview, a section on running ROOT containers with terminal code snippets for ROOT6 and ROOT5, and a section on sharing files with the container. The page is part of a website with a blue header and navigation tabs.

Running ROOT in Containers

Overview

This page was created as a quick reference for running different versions of ROOT in Docker containers. For technical details The PHENIX Collaboration published a small selection of Docker images on [Docker Hub](#) for everyone's use. For that reason, identical.

Running ROOT Containers

The following assumes that the user has installed Docker on their machine.

```
# To run the current version of ROOT - which is ROOT6 at the time of writing:
docker run -it --ipc=host --rm -v /tmp/.X11-unix:/tmp/.X11-unix rootproject/root

# To run ROOT5 preserved as a Docker image:
docker run -it --ipc=host --rm -v /tmp/.X11-unix:/tmp/.X11-unix phenixcollaboration/tools:s17_root5
```

Since typing a long command like this may be tedious and error-prone one can easily define an alias in any Linux-like environment

```
# create the alias which works for any X11-capable container:
alias droot="docker run -it --ipc=host --rm -v /tmp/.X11-unix:/tmp/.X11-unix '

# Run the ROOT5 container based on the PHENIX image
droot phenixcollaboration/tools:s17_root5
```

Of course the whole command including the name of the image can be aliased just as easily. Creating shell functions and/or one can conveniently define *different* volumes to be mounted, as necessary - see the section below.

Sharing files with the container

In order to share the filesystem between the host machine and the Docker container one needs to use the "-v" option as explained (populated is with data on the local host) it will be accessible in the ROOT5 container as the "user" folder if the following command

```
docker run -it --ipc=host --rm -v /tmp/.X11-unix:/tmp/.X11-unix -v myvolume:/user phenixcollaboration/
```

If a command such as this one is used repeatedly it might be a good idea to create aliases similar to one described in the previous

Shell access

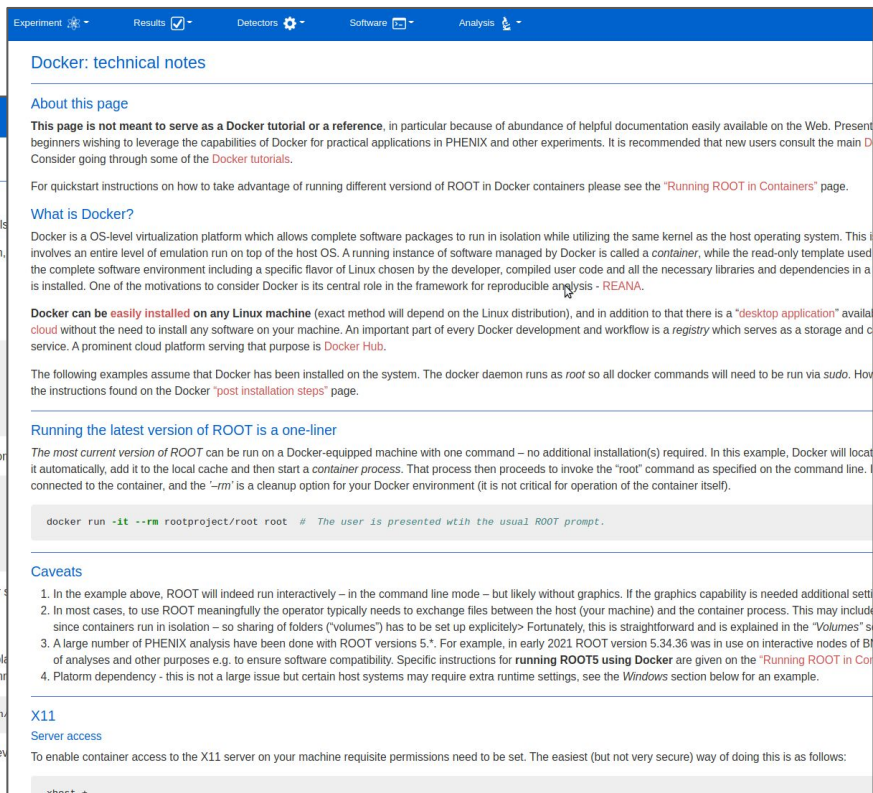
In order to gain shell access to the container instead of jumping into ROOT one needs to simply add "bash" to the end of the command line

```
docker run -it --ipc=host --rm -v /tmp/.X11-unix:/tmp/.X11-unix -v myvolume:/user phenixcollaboration/tools:s17
```

This will start the bash shell which will function normally. This may be necessary to check presence of requisite files, macros and other settings

Customizing your ROOT container

Please see the ["technical notes"](#) page for advice how to customize the Docker image of ROOT5 with user-defined folders etc.



The screenshot shows the 'Docker: technical notes' page. It includes an 'About this page' section, a 'What is Docker?' section, and a 'Running the latest version of ROOT is a one-liner' section. The page is part of a website with a blue header and navigation tabs.

Docker: technical notes

About this page

This page is not meant to serve as a Docker tutorial or a reference, in particular because of abundance of helpful documentation easily available on the Web. Present beginners wishing to leverage the capabilities of Docker for practical applications in PHENIX and other experiments. It is recommended that new users consult the main [D](#). Consider going through some of the [Docker tutorials](#).

For quickstart instructions on how to take advantage of running different versions of ROOT in Docker containers please see the ["Running ROOT in Containers"](#) page.

What is Docker?

Docker is an OS-level virtualization platform which allows complete software packages to run in isolation while utilizing the same kernel as the host operating system. This involves an entire level of emulation run on top of the host OS. A running instance of software managed by Docker is called a *container*, while the read-only templates used to create the complete software environment including a specific flavor of Linux chosen by the developer, compiled user code and all the necessary libraries and dependencies in a is installed. One of the motivations to consider Docker is its central role in the framework for reproducible analysis - [REANA](#).

Docker can be easily installed on any Linux machine (exact method will depend on the Linux distribution), and in addition to that there is a ["desktop application"](#) available cloud without the need to install any software on your machine. An important part of every Docker development and workflow is a registry which serves as a storage and distribution service. A prominent cloud platform serving that purpose is [Docker Hub](#).

The following examples assume that Docker has been installed on the system. The docker daemon runs as root so all docker commands will need to be run via `sudo`. How the instructions found on the Docker ["post installation steps"](#) page.

Running the latest version of ROOT is a one-liner

The most current version of ROOT can be run on a Docker-equipped machine with one command – no additional installation(s) required. In this example, Docker will locate it automatically, add it to the local cache and then start a *container* process. That process then proceeds to invoke the "root" command as specified on the command line, it connected to the container, and the "-rm" is a cleanup option for your Docker environment (it is not critical for operation of the container itself).

```
docker run -it --rm rootproject/root root # The user is presented with the usual ROOT prompt.
```

Caveats

1. In the example above, ROOT will indeed run interactively – in the command line mode – but likely without graphics. If the graphics capability is needed additional settings
2. In most cases, to use ROOT meaningfully the operator typically needs to exchange files between the host (your machine) and the container process. This may include since containers run in isolation – so sharing of folders ("volumes") has to be set up explicitly. Fortunately, this is straightforward and is explained in the "Volumes" section
3. A large number of PHENIX analysis have been done with ROOT versions 5.*. For example, in early 2021 ROOT version 5.34.36 was in use on interactive nodes of BNL for analyses and other purposes e.g. to ensure software compatibility. Specific instructions for [running ROOT5 using Docker](#) are given on the ["Running ROOT in Containers"](#) page
4. Platform dependency - this is not a large issue but certain host systems may require extra runtime settings, see the [Windows](#) section below for an example.

X11

Server access

To enable container access to the X11 server on your machine requisite permissions need to be set. The easiest (but not very secure) way of doing this is as follows:

```
sudo xhost +
```

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