Data & Analysis Preservation: status update

Maxim Potekhin
Nuclear and Particle Physics Software Group

Brookhaven National Laboratory

PHENIX DAP Meeting
08/05/2021

Overview

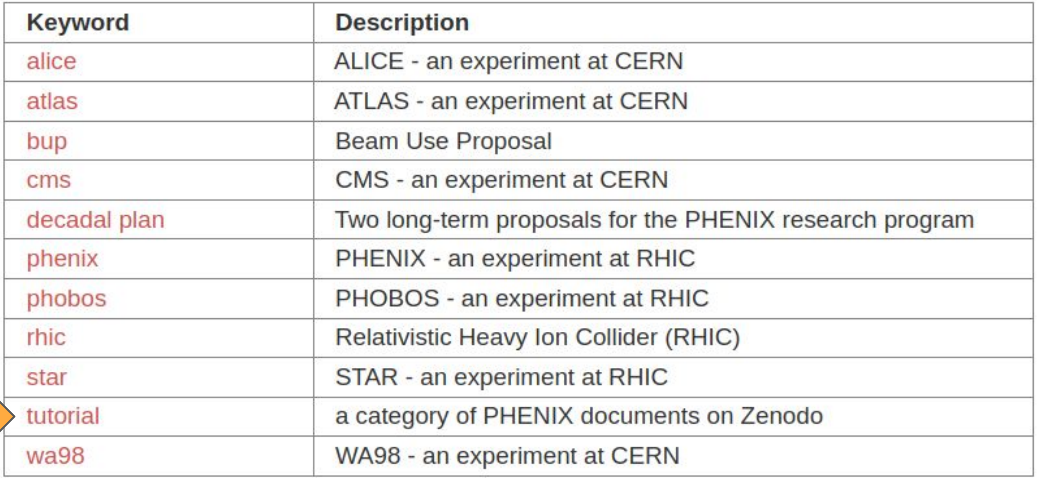
- HEPData
- Website
- Data archive/URLs (pre-HEPData)
- REANA

HEPData

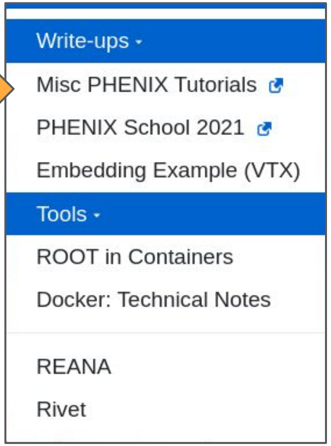
- Master spreadsheet updated
- https://docs.google.com/spreadsheets/d/1rABxzuM-h9Rukz08ut_m8xnMo0B_J1LKre8bM7B7264/edit?usp=sharing
- PPG081*, PPG139, PPG209 finalized
- Some other active items in the pipeline

Website

- The “tutorial” keyword introduced and added to existing items
- Separate “tutorials” page eliminated as it’s no longer needed, for same reason
- More School materials uploaded



Keyword	Description
alice	ALICE - an experiment at CERN
atlas	ATLAS - an experiment at CERN
bup	Beam Use Proposal
cms	CMS - an experiment at CERN
decadal plan	Two long-term proposals for the PHENIX research program
phenix	PHENIX - an experiment at RHIC
phobos	PHOBOS - an experiment at RHIC
rhic	Relativistic Heavy Ion Collider (RHIC)
star	STAR - an experiment at RHIC
tutorial	a category of PHENIX documents on Zenodo
wa98	WA98 - an experiment at CERN



Website - group pages?

- An idea - current working groups can (should?) have presence on the site.
- That would help preserve some of the current knowledge and perhaps useful links and documents.

Legacy publication data content (pre-HEPData)

- Comments from the EC - old links provided along with PHENIX publications to publishing houses are broken (due to migration to the new site)
- Fixing all these on the publishers’ side is not practical
- Currently all data is still available on the internal website - but not externally
- The layout of the directory is “ad hoc”, there is no standard structure or convention
- Some entries are links to folders elsewhere (/p etc), some links broken
- Contents (after link fixes) are now captured in our repo
  https://github.com/PhenixCollaboration/documentation/tree/master/assets/data
- Deployment? Next slide...

Legacy publication data: where to host?

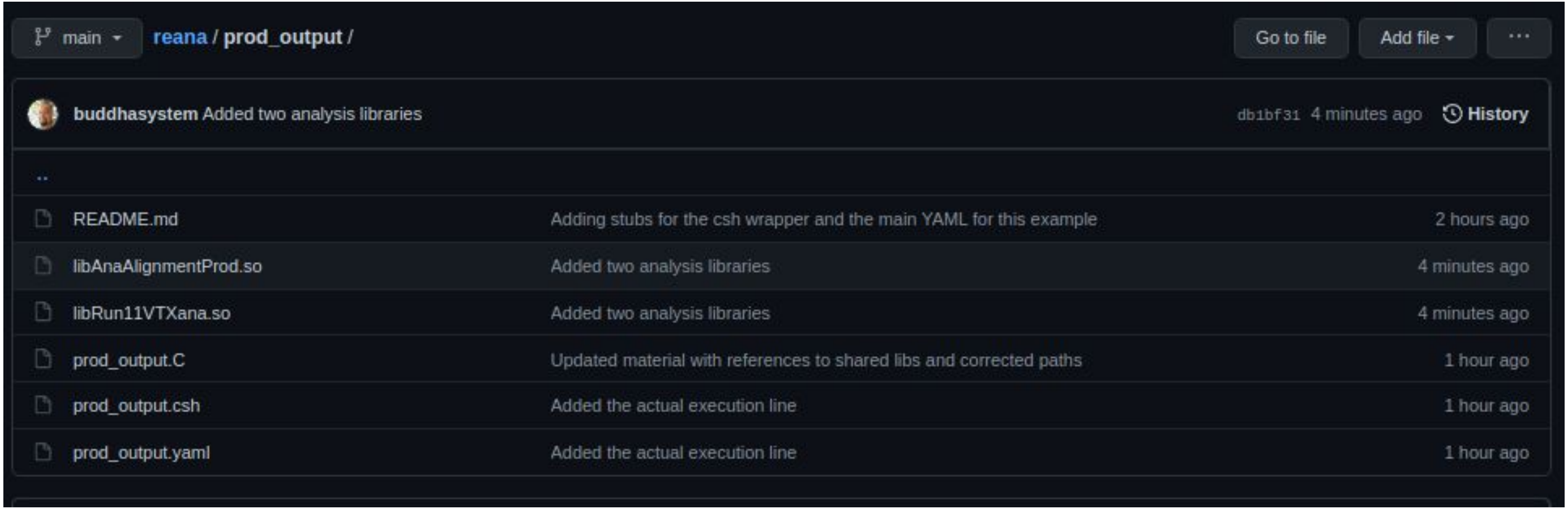
- A logical place would be on the Apache server, in the same way as hosting users' directories (i.e. adding a separate directory tree to the configuration, distinct from the site proper)
- Not approved by Chris and Mizuki on the grounds that it would be hard to assign blame in case of a security incident (however unlikely this is, with the content that's 100% static by definition)
- Adding these data to the new site is awkward but should be doable - it would then be in the same repository as the code for the site itself

REANA

- DB access problem solved, as reported last time
- As of 3 weeks ago, a successful REANA demonstration using an image provided by SDCC, along with the PHENIX software deployed to CVMFS
  - pisa
  - VTX macro - worked, some comments from Gabor about how distributions look
- Materials are on GitHub
- Not much new activity, standing by to work with users/analysis people to implement analyses in REANA
- Would be a nice complement for the possible next submission to OpenData (we can upload some ROOT files for preservation and annotate them)

REANA: run16 QA example on GitHub (redux)

https://github.com/PhenixCollaboration/reana/tree/main/prod_output



buddhasystem Added two analysis libraries — db1bf31 4 minutes ago

..

README.md	Adding stubs for the csh wrapper and the main YAML for this example	2 hours ago
libAnaAlignmentProd.so	Added two analysis libraries	4 minutes ago
libRun11VTXana.so	Added two analysis libraries	4 minutes ago
prod_output.C	Updated material with references to shared libs and corrected paths	1 hour ago
prod_output.csh	Added the actual execution line	1 hour ago
prod_output.yaml	Added the actual execution line	1 hour ago

REANA submission for run16 QA (redux)

```
version: 0.0.1
inputs:
  files:
    - /afs/rhic.bnl.gov/phenix/etc/odbc.ini
    - ./prod_output.csh
    - ./prod_output.C
    - ./libAnaAlignmentProd.so # any analysis-specific shared libraries can be organized optimally as folders etc
    - ./libRun11VTXana.so
    - /phenix/crs/agg/run16/run16AuAu_200GeV_CA_pro111_agg/CNT_MB/CNT_MB_run16AuAu_200GeV_CA_pro111-0000459208-9000.root
    - /phenix/crs/agg/run16/run16AuAu_200GeV_CA_pro111_agg/DST_SVX_MB/DST_SVX_MB_run16AuAu_200GeV_CA_pro111-0000459208-9000.root
    - /phenix/crs/agg/run16/run16AuAu_200GeV_CA_pro111_agg/DST_EVE_MB/DST_EVE_MB_run16AuAu_200GeV_CA_pro111-0000459208-9000.root
workflow:
  type: serial
  specification:
    steps:
      - environment: 'registry.sdcc.bnl.gov/sdcc-fabric/rhic_sl7_ext:1.3'
        commands:
        - chmod +x ./prod_output.csh
        - mv ./phenix/crs/agg/run16/run16AuAu_200GeV_CA_pro111_agg/*/*.root .
        - rm -fr ./phenix
        - ls > output.txt
        - ./prod_output.csh >> output.txt
outputs:
  files:
    - output.txt
    - test_459208.root
```

REANA roadmap (redux)

- There has been a substantial investment of effort invested in the images and REANA
- Conservatively 80% of work on tuning Docker images has been done
- Simple - but realistic - PHENIX examples work
- To start seeing dividends of this work we need wider community engagement - really the key to completing this work area and making it meaningful
- Should we organize a REANA seminar for the PHENIX community, as a follow up to the School tutorials?
- Can we identify ~2 PPGs who can be asked to “reanify” some parts of their analysis?
  - There are some direct benefits for the groups (self-documented reproducible procedures)
  - REANA take-up can be facilitated by providing templates for submission files