Data & Analysis Preservation: status update

Maxim Potekhin

Nuclear and Particle Physics Software Group



PHENIX DAP Meeting 09/16/2021



Overview

- HEPData
- Website
- REANA
- New tools
- Meeting in Sept





HEPData

- Master spreadsheet updated
- https://docs.google.com/spreadsheets/d/1rABxzuM-h9Rukz08ut_m8xnMo0B_J1LKre8bM7B7264/edit?usp=sharing
- PPG147 needs to be reviewed and corrected (talking to Takahito)
- PPG238 waiting to be approved by Sanghwa
- Some other active items in the pipeline (003, 023, 212, 238 etc)
- Work on 202 to commence in September (which is now)





Website

- Landing page cleaned up, added REANA to the text in the banner
- Removal of material no longer relevant (e.g. local catalog of papers)
- Better formatting in the keyword table
- Added ~8 new conferences (thanks Gabor) finalized 2018
 - A few more coming
- Migration of legacy pre-HEPData numerical data (next slide)





Migration of legacy pre-HEPData material

- Requested by the EC in July 2021
 - Motivation links provided to publishers etc became broken since the website migration
- Step 1: capture material on GitHub (text file, some plots)
 - NB. some file locations are links and/or non-standard, so not a straight copy
- Step 2: manual checks and edit in some cases
- Step 3: add content (mostly text files but also some PDF and images) to the website, create data structures to manage that, create a dedicated page
- Important ensure legacy links still work





Top-cited

- It was decided that it's helpful to have a list of top-cited papers on the website
- Direct links to InspireHEP and HEPData where available
- Some maintenance will be required





Data catalog: both legacy and HEPData

- # The file: data/publications.md
- # https://raw.githubusercontent.com/PhenixCollaboration/web/master/ data/publications.yml
- # Example of entries:
- title: 'Creation of guark-gluon plasma droplets with three distinct geometries'

inspire: 1672133

data: legacy

legacyURL: ppg216 data.html

- title: 'Dihadron azimuthal correlations in Au+Au collisions at √s_{NN}=200GeV'

inspire: 778396 data: hepdata

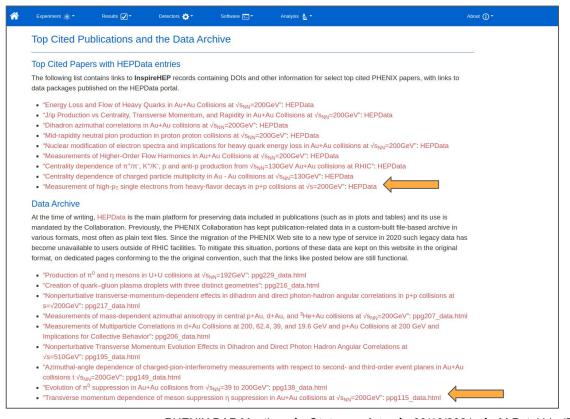
Created a macro which generates proper links automatically for HEPData, Inspire and legacy content where needed

The page on the next slide is automatically generated



The new publications page









Legacy links

- Two types of links were communicated to the publishers in the past
 - To the umbrella publication page "papers.html"
 - To individual pages like "ppg115_data.html"
- In the new setup, both types of legacy links work provided the numeric content for a particular paper has been migrated
 - A semi-manual process but can be done quickly as requested
 - See many items on the previous slide that have been migrated
 - The new "papers" page contains a link to that one



REANA



- As previously discussed with SDCC increase the storage for the REANA cluster
 - Took some time because of coordinating with other users
- Storage increase: 6TB was the original target and coordination took some time
 - Storage was added on Sept.15th
 - Will need to be tested with our REANA jobs
 - ...but apparently due to RAID configuration (redundancy etc) we actually get only 2.6TB
 - Still more than 10 times larger than before, opens possibilities for analysis
 - If we want more storage, we shall need justification i.e. real analyses running
 - NB. Storage shared across the cluster, may require coordination with other collaborations
 - Learning about Kubernetes deployment by sPHENIX and how it meshes with REANA
- As mentioned previously, uptake of REANA in PHENIX is still on the to-do list, pending capturing workflow examples (cf. Niv's effort)



Misc

- OpenData
- REANA
- Given the steady state, should we meet again in a month?

