

# Data & Analysis Preservation: status update

Maxim Potekhin

*Nuclear and Particle Physics Software Group*



**Brookhaven™**  
National Laboratory

***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](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 quark–gluon plasma droplets with three distinct geometries'
```

```
  inspire: 1672133
```

```
  data: legacy
```

```
  legacyURL: ppg216_data.html
```

```
- title: 'Dihadron azimuthal correlations in Au+Au collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ '
```

```
  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



[Experiment](#) [Results](#) [Detectors](#) [Software](#) [Analysis](#) [About](#)

## Top Cited Publications and the Data Archive

### Top Cited Papers with HEPData entries

The following list contains links to **InspireHEP** records containing DOIs and other information for select top cited PHENIX papers, with links to data packages published on the HEPData portal.

- "Energy Loss and Flow of Heavy Quarks in Au+Au Collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": HEPData
- "J/ $\psi$  Production vs Centrality, Transverse Momentum, and Rapidity in Au+Au Collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": HEPData
- "Dihadron azimuthal correlations in Au+Au collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": HEPData
- "Mid-rapidity neutral pion production in proton proton collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": HEPData
- "Nuclear modification of electron spectra and implications for heavy quark energy loss in Au+Au collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": HEPData
- "Measurements of Higher-Order Flow Harmonics in Au+Au Collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": HEPData
- "Centrality dependence of  $\pi^0/\pi^+$ ,  $K^0/K^+$ , p and anti-p production from  $\sqrt{s_{NN}}=130\text{GeV}$  Au+Au collisions at RHIC": HEPData
- "Centrality dependence of charged particle multiplicity in Au - Au collisions at  $\sqrt{s_{NN}}=130\text{GeV}$ ": HEPData
- "Measurement of high- $p_T$  single electrons from heavy-flavor decays in p+p collisions at  $\sqrt{s}=200\text{GeV}$ ": HEPData

### Data Archive

At the time of writing, **HEPData** is the main platform for preserving data included in publications (such as in plots and tables) and its use is mandated by the Collaboration. Previously, the PHENIX Collaboration has kept publication-related data in a custom-built file-based archive in various formats, most often as plain text files. Since the migration of the PHENIX Web site to a new type of service in 2020 such legacy data has become unavailable to users outside of RHIC facilities. To mitigate this situation, portions of these data are kept on this website in the original format, on dedicated pages conforming to the the original convention, such that the links like posted below are still functional.

- "Production of  $\pi^0$  and  $\eta$  mesons in U+U collisions at  $\sqrt{s_{NN}}=192\text{GeV}$ ": ppg229\_data.html
- "Creation of quark-gluon plasma droplets with three distinct geometries": ppg216\_data.html
- "Nonperturbative transverse-momentum-dependent effects in dihadron and direct photon-hadron angular correlations in p+p collisions at  $\sqrt{s}=200\text{GeV}$ ": ppg217\_data.html
- "Measurements of mass-dependent azimuthal anisotropy in central p+Au, d+Au, and  $^3\text{He}$ +Au collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": ppg207\_data.html
- "Measurements of Multiparticle Correlations in d+Au Collisions at 200, 62.4, 39, and 19.6 GeV and p+Au Collisions at 200 GeV and Implications for Collective Behavior": ppg206\_data.html
- "Nonperturbative Transverse Momentum Evolution Effects in Dihadron and Direct Photon Hadron Angular Correlations at  $\sqrt{s}=510\text{GeV}$ ": ppg195\_data.html
- "Azimuthal-angle dependence of charged-pion-interferometry measurements with respect to second- and third-order event planes in Au+Au collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": ppg149\_data.html
- "Evolution of  $\pi^0$  suppression in Au+Au collisions from  $\sqrt{s_{NN}}=39$  to  $200\text{GeV}$ ": ppg138\_data.html
- "Transverse momentum dependence of meson suppression  $\eta$  suppression in Au+Au collisions at  $\sqrt{s_{NN}}=200\text{GeV}$ ": ppg115\_data.html



# 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

- 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. 15<sup>th</sup>
  - 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?