

# Data & Analysis Preservation: status update

Maxim Potekhin

*Nuclear and Particle Physics Software Group*



**Brookhaven**<sup>™</sup>  
National Laboratory

***PHENIX DAP Meeting***

11/18/2021

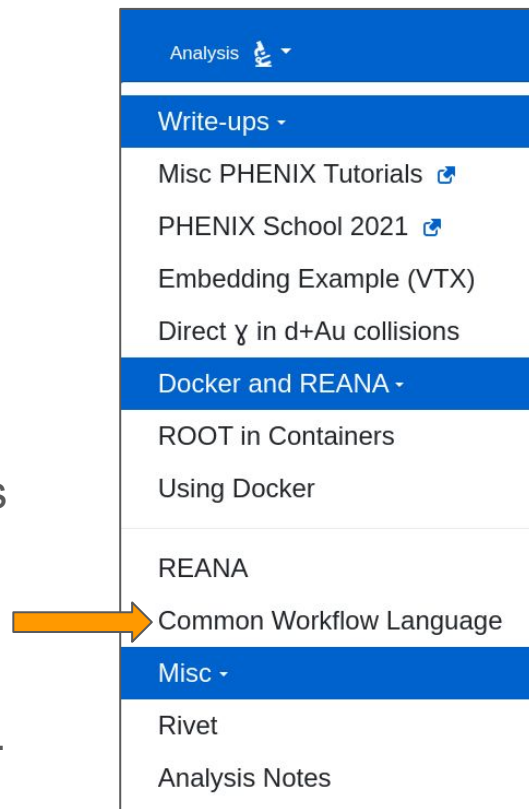


# Overview

- HEPData
  - Steady state activity
  - pg202 published
  - pg212 approaching final stage
  - Some items went dormant (people busy with other work)
  - See the spreadsheet for details:
  - [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)
- Website (in the following few slides)
- REANA - Common Workflow Language (CWL)
- Niv's note impressions

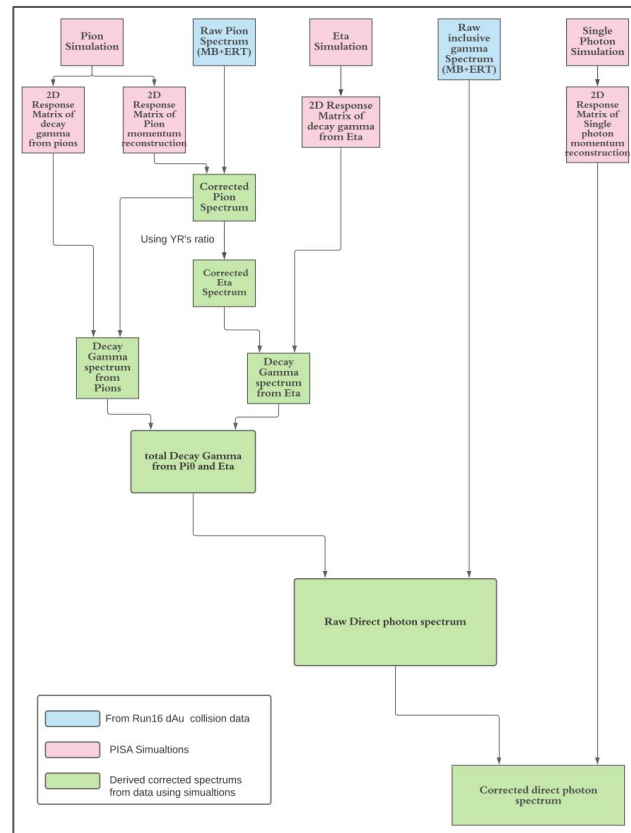
# Website

- Improved the layout of a few pages, general cleanup
- Added a CWL page (needed to document parallel workflows in REANA) - next slide
- Modified the “Analysis” drop-down menu to improve readability
- Added content to the “Direct gamma” page based on Niv’s analysis note
- ...also the analysis diagram (next slide)
- Still ways to go to become a reference point for analysis... Contributions needed

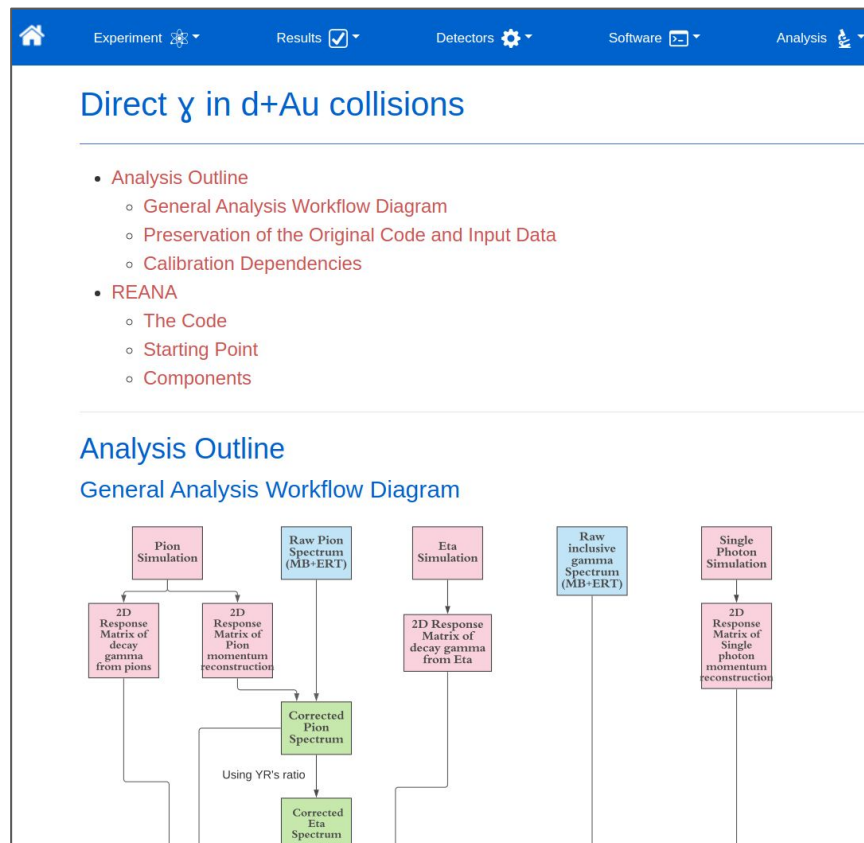


# Website - direct gamma analysis diagram added

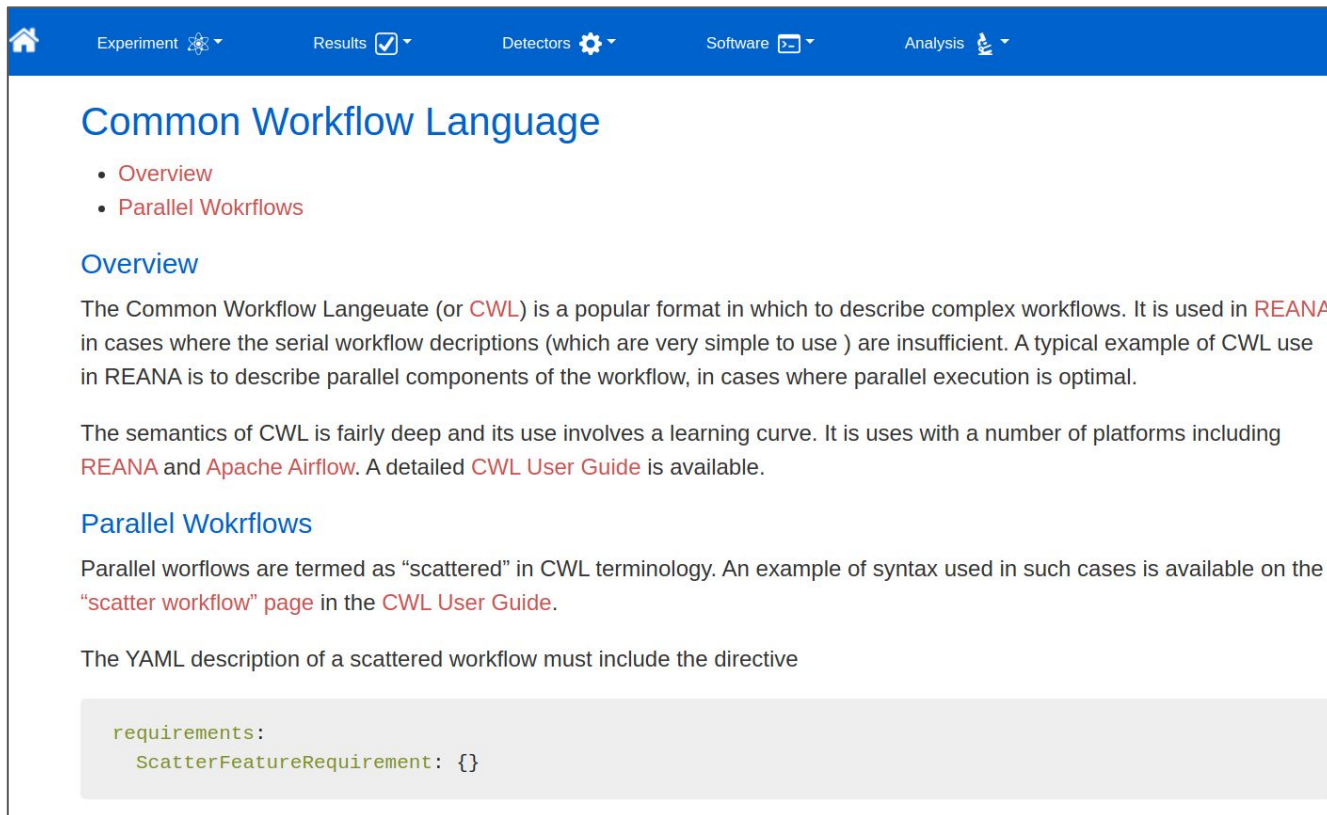
- Documentation work will continue
- As discussed before, some written material here would be extremely helpful both for the ongoing REANA work and also for actual future use
- REANA or plain scripting preservation won't work unless there is a narrative for it - also important for longer term
- Can we get a hold on a few original plots from the analysis note to illustrate crucial steps? (Should obviously be blessed)



# Website - the updated direct gamma page



# Website - the new CWL page

A screenshot of the Common Workflow Language website. The top navigation bar is blue with white text and icons for Experiment, Results, Detectors, Software, and Analysis. The main content area has a white background. The title "Common Workflow Language" is in blue. Below it are two red links: "Overview" and "Parallel Wokflows". The "Overview" section has a blue heading and a paragraph explaining CWL. The "Parallel Wokflows" section has a blue heading and a paragraph explaining scattered workflows. A code block at the bottom shows a YAML snippet for requirements.

## Common Workflow Language

- [Overview](#)
- [Parallel Wokflows](#)

### Overview

The Common Workflow Langeuate (or [CWL](#)) is a popular format in which to describe complex workflows. It is used in [REANA](#) in cases where the serial workflow decriptions (which are very simple to use ) are insufficient. A typical example of CWL use in REANA is to describe parallel components of the workflow, in cases where parallel execution is optimal.

The semantics of CWL is fairly deep and its use involves a learning curve. It is uses with a number of platforms including [REANA](#) and [Apache Airflow](#). A detailed [CWL User Guide](#) is available.

### Parallel Wokflows

Parallel worflows are termed as “scattered” in CWL terminology. An example of syntax used in such cases is available on the “[scatter workflow](#)” [page](#) in the [CWL User Guide](#).

The YAML description of a scattered workflow must include the directive

```
requirements:
  ScatterFeatureRequirement: {}
```

# Niv's analysis note

- A rather involved calibrations component not included in the diagram and notes, so these will have to be “prior” at least for now
- Some simulation components (not a part of the workflow in the “canonical” diagram) has dependencies like  
`/plhf/plhf3/nnovitzk/mazsi_Test/ccnt/source/emc-evaluation` i.e. outside of the preserved code base
- In the big picture, how much of this analysis infrastructure do we want to keep?

# REANA - Parallel Workflows - CWL



- Completed “scattered workflow” tutorials: [http://www.commonwl.org/user\\_guide/](http://www.commonwl.org/user_guide/)
- Created interactive testing environment on SDCC nodes
- Starting work on implementing ROOT macros in this format
- There are still features in Niv’s code from interactive/Condor-based analysis e.g. symbolic links to specific locations, general folder structure/assumptions, scripting conventions targeted at Condor etc - some re-engineering is required, needs work