

Facilitating future open data reuse via continuous integration of actionable data analysis examples

Marco Donadoni, Audrius Mečionis, Giuseppe Steduto, Tibor Šimko

CERN

PV2023 conference, Geneva, Switzerland, 2-4 May 2023

<https://indico.cern.ch/event/1188041/contributions/5309206/>

CERN Open Data

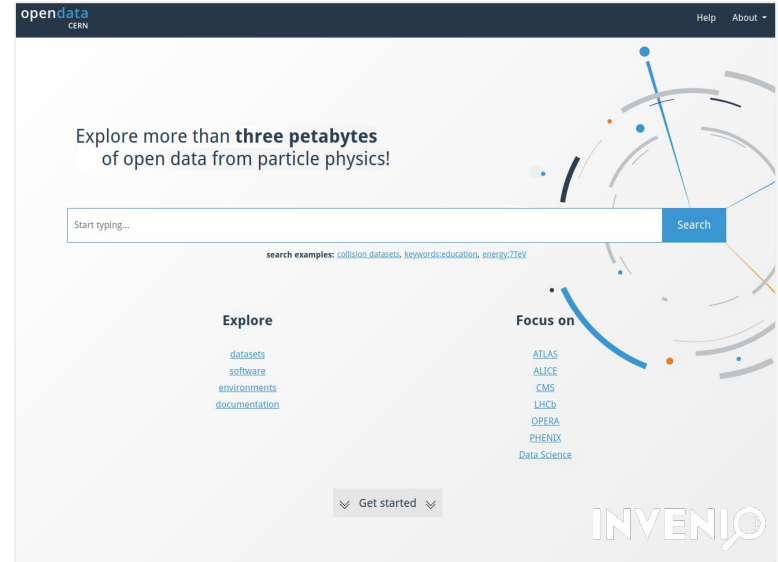
Digital repository for event-level particle physics open data

- collision and simulated datasets for research
- derived datasets for education
- configuration files and documentation
- virtual machines and container images
- software tools and analysis example

Launched in November 2014

Current size (April 2023)

- over 15 000 bibliographic records
- over 1 500 000 files
- over 3 petabytes

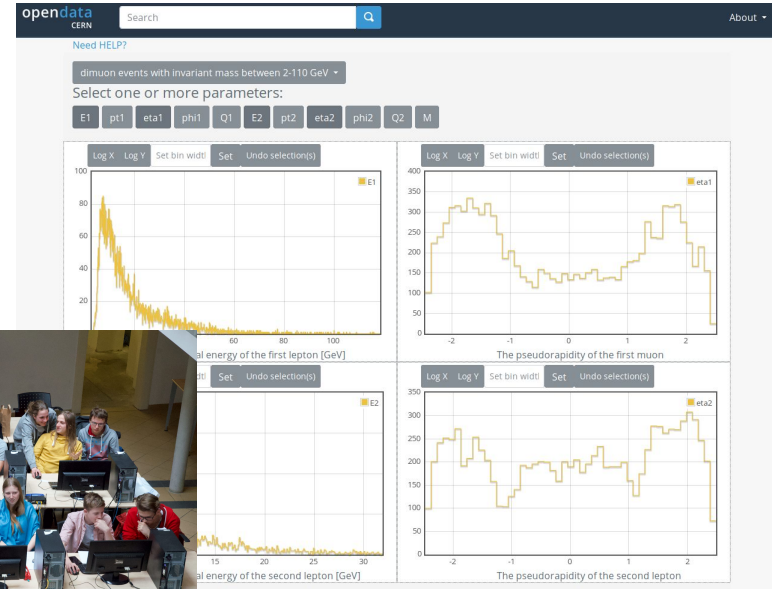
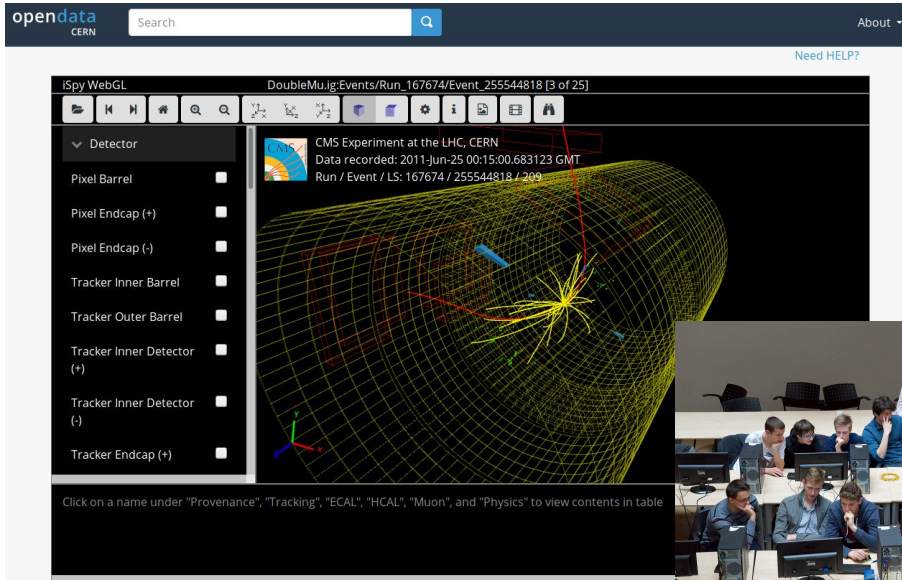


<https://opendata.cern.ch>

Developed by CERN in close collaboration with LHC (and non-LHC!) experiments



Education-oriented use cases



Interactive collision event displays and basic histogramming from derived datasets

Research-oriented use cases

Content metadata

Provenance metadata

Dataset characteristics
40926332 events, 3943 files, 14.2 TB in total.

System details
Recommended global tag for analysis: FT53_V21A_AN6
Recommended release for analysis: CMSW_5_3_32

How were these data selected?
Events stored in this primary dataset were selected because of the presence two or more high-energy jets a b-quark-tag requirement in the event.

Data taking / HLT
The collision data were assigned to different RAW datasets using the following HLT configuration.

Data processing / RECO
This primary AOD dataset was processed from the RAW dataset by the following step:
Step: RECO
Release: CMSW_5_3_7_patch5
Global tag: FT_R_53_V18-All
Configuration file for RECO step reco_2012D_BjetPlusX

HLT trigger paths
The possible HLT trigger paths in this dataset are:
HLT_Dijet0Ea2q6_BTagP3DFastPV
HLT_Dijet0Ea2q6_BTagP3DFastPVLoose
HLT_DIPFjet80_DIPFjet30_BTagCSVd7d05
HLT_DIPFjet80_DIPFjet30_BTagCSVd7d05d03
HLT_DIPFjet80_DIPFjet30_BTagCSVd7d05d03_PFDJet120
HLT_DIPFjet80_DIPFjet30_BTagCSVd7d05d05
HLT_jet160Ea2q6_jet120Ea2q6_DiBTagP3DFastPVLoose

Run virtual machines and containers with the same physics environment



CMS open data workshops for research use

Enables independent research

Explore research-grade primary datasets

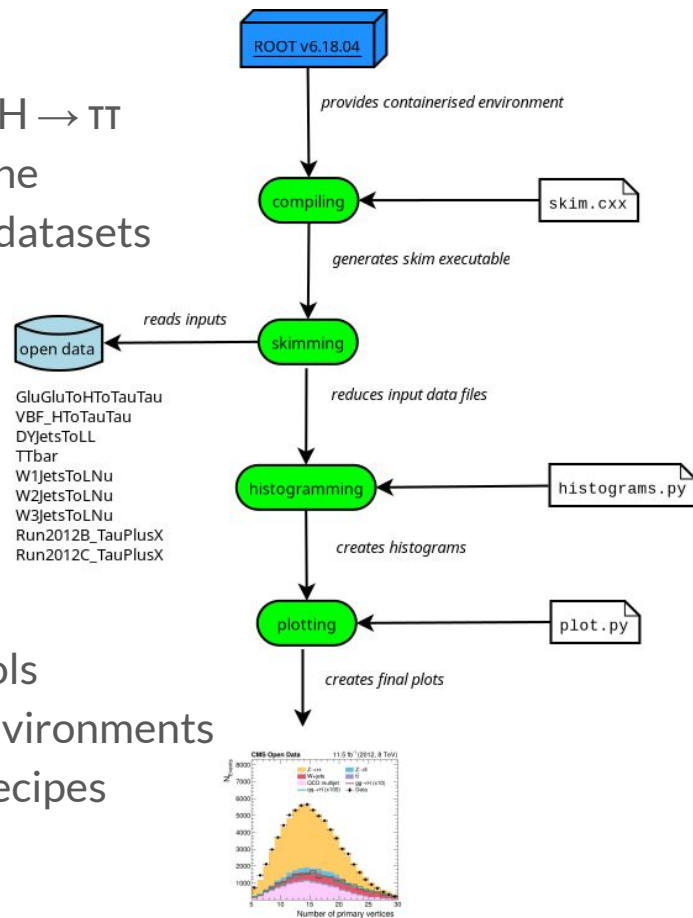
Enables independent research 4

How to use the data?

The screenshot shows the OpenData portal interface. The top section displays the dataset title "W3jetsToLNu dataset in reduced NanoAOD format for education and outreach" by Wunsch, Stefan. Below this, there are sections for "Description", "Use with", "Related datasets", and "Dataset characteristics". The "Description" section includes a detailed analysis of Higgs boson decays to two tau leptons, mentioning the use of data and simulation of events at the CMS detector from 2012. Two example plots are shown at the bottom, comparing data points with various simulation models for the visible di-tau mass and the number of primary vertices.

Data is accompanied with analysis examples

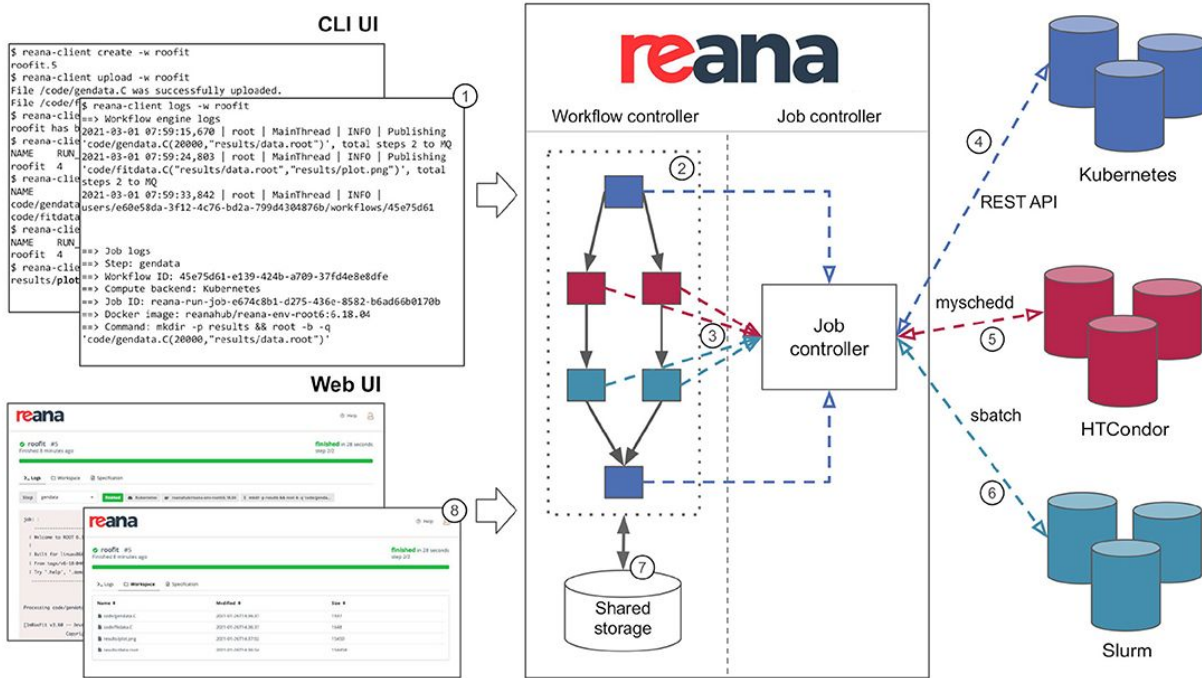
An example studying $H \rightarrow \tau\tau$ lepton decays uses nine published CMS open datasets



Data is not enough!

- software and tools
- containerised environments
- computational recipes

REANA reusable analyses



Multiple compute backends:

- Kubernetes
- HTCondor
- Slurm

Different workflow languages:

- CWL
- Serial
- Snakemake
- Yadage

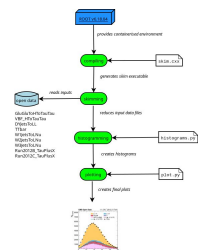
Multiple means of use:

- Command-line client
- Web UI

<https://www.reana.io>

Run containerised analysis workflows on the cloud

H \rightarrow $\tau\tau$ example running on REANA



```

1 version: 0.6.0
2 inputs:
3 files:
4   - skim.cxx
5   - histograms.py
6   - plot.py
7 workflow:
8 type: serial
9 specification:
10 steps:
11   - name: compiling
12     environment: reanahub/reana-env-root6:6.18.04
13     commands:
14       - g++ -g -O3 -Wall -Wextra -Wpedantic -o skim skim.cxx `root-config --cflags --libs`
15   - name: skimming
16     environment: reanahub/reana-env-root6:6.18.04
17     commands:
18       - ./skim
19   - name: histogramming
20     environment: reanahub/reana-env-root6:6.18.04
21     commands:
22       - python histograms.py
23   - name: plotting
24     environment: reanahub/reana-env-root6:6.18.04
25     commands:
26       - python plot.py
    
```

Run example on REANA

Consult workflow logs

Structure data usage workflow

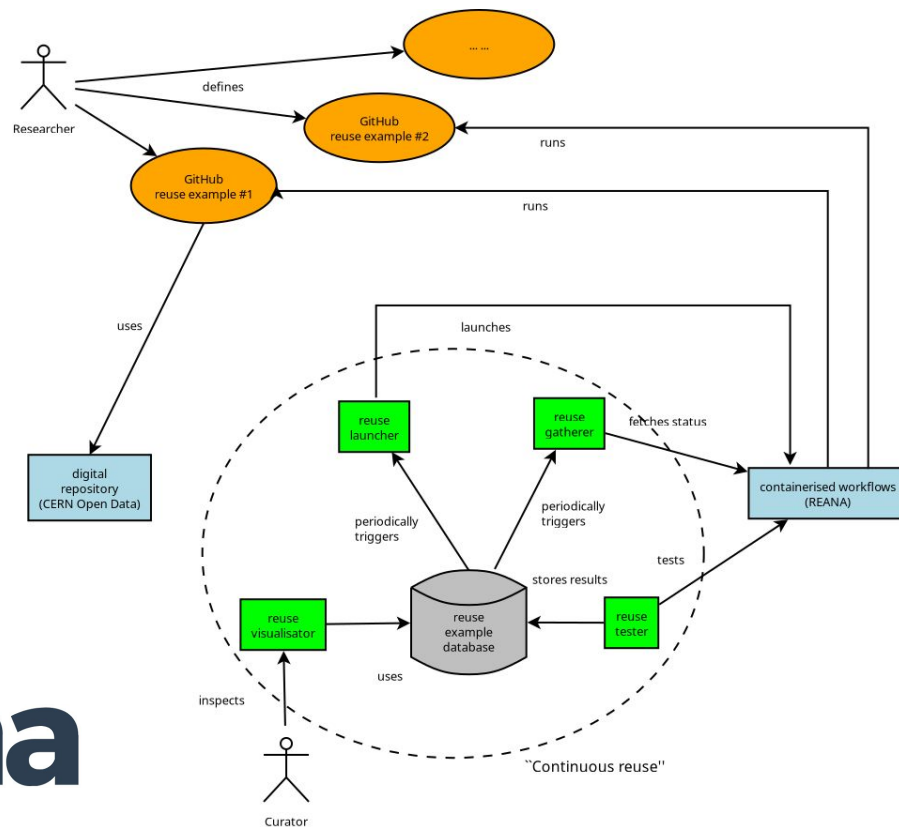
Visualise workflow outputs 7

Continuous reuse

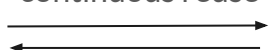
Continuous reuse idea:

- run data reuse examples periodically
- collect data reuse run information
- test data reuse run outputs
- visualise statistics on a web dashboard

Ensures the accessibility of data and correctness of data reuse examples



“continuous reuse”



H → TT reuse example tests

Using Gherkin feature files allows data curators to express desired tests in domain-oriented natural language.

Developed tests allowing to check for:

- workspace content
- workspace size
- file sizes
- file checksums
- log content
- job runtime durations

```
Feature: cms-htautau-nanoaod

Scenario: Workspace content
  When the workflow is finished
  Then the workspace should contain "njets.png"
  And the workspace should contain "phi_1.png"
  And the workspace should contain "phi_2.png"
  And the workspace should contain "pt_1.png"
  And the workspace should contain "jpt_2.png"
  And the workspace should contain "jdeta.png"

Scenario: Workspace size
  When the workflow is finished
  Then the workspace size should be less than 75 MiB

Scenario: Log content
  When the workflow is finished
  Then the job logs of the step "skimming" should contain "Event has good muons: pass=36921"
  And the job logs of the step "skimming" should contain "Event has good taus: pass=38041"
  And the job logs of the step "histogramming" should contain "Muon transverse mass cut for
  W+jets suppression: pass=5063"
  And the job logs of the step "plotting" should contain "pdf file m_1.pdf has been created"
  And the job logs of the step "plotting" should contain "pdf file m_2.pdf has been created"

Scenario: Run duration
  When the workflow is finished
  Then the workflow run duration should be less than 25 minutes
  And the duration of the step "compiling" should be less than 1 minutes
  And the duration of the step "skimming" should be less than 20 minutes
  And the duration of the step "histogramming" should be less than 3 minutes
  And the duration of the step "plotting" should be less than 3 minutes
```

Example of a Gherkin file defining expected outcomes of the H → TT analysis reuse run

User dashboard: Home page

Grafana dashboard allows to visualise the collected data

- displays a history of various reuse examples and their statistics
- allows to quickly check the last success and failure timestamps
- shows the results of last five runs



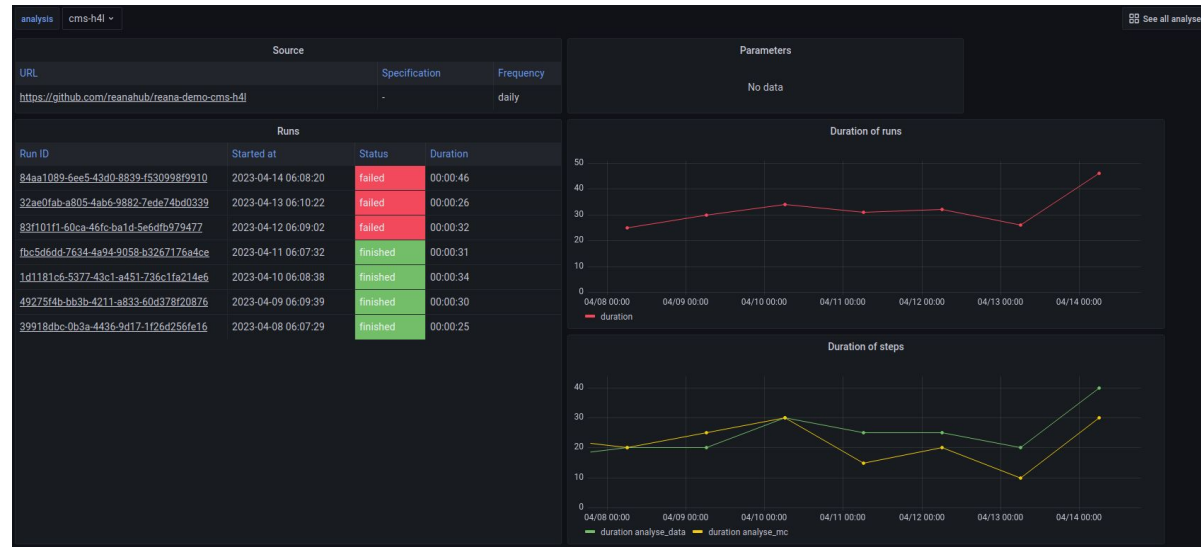
Name ↑	Last success	Last failure	Last duration	R1	R2	R3	R4	R5
alice-lego-train-test-run	5 hours ago		00:01:35	Green	Green	Green	Green	Green
alice-pt-analysis	5 hours ago		00:01:10	Green	Green	Green	Green	Green
atlas-recast	5 hours ago		00:01:10	Green	Green	Green	Green	Green
cms-dimuon-mass-spectrum	8 days ago		00:01:03	Green	Green	Green	Green	Green
cms-dimuon-spectrum	5 hours ago		00:01:22	Green	Green	Green	Green	Green
cms-dimuon-spectrum-nanoaod	5 hours ago		00:01:52	Green	Green	Green	Green	Green
cms-h4l	5 hours ago	2 days ago	00:02:02	Green	Green	Red	Red	Red
cms-h4l-nanoaod	5 hours ago		00:03:46	Green	Green	Green	Green	Green
cms-htautau-nanoaod	5 hours ago		00:07:08	Green	Green	Green	Green	Green

Continuous reuse dashboard home page showing the overall status of reuse analyses

User dashboard: Detailed view

Another dashboard view offers detailed information about one particular data reuse example

- shows the overall success/failure statistics over time
- displays the duration of runs and individual steps in easily readable charts
- enables early detection of performance issues



The detailed page of the dashboard for one particular reuse example

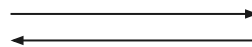
Conclusions

- Adding value to preserved data through actionable containerised workflows!
 - containers allow to encapsulate the original computing environment around the data
 - data production workflows allow to verify the data provenance information
 - data usage workflows allow to understand the data reuse through concrete examples
- “Continuous reuse” helps to discover problems early
 - accessibility and readability of data across time
 - validity of data usage examples across time
 - discover side issues due to changing versions of software protocols etc
- Actionable data usage examples help to pass the data knowledge to future generations
- Data + Code + Environments + Workflows = Reusable Knowledge

"adaptable software examples [are] the most efficient way to pass on the knowledge needed for research-level studies on these data" — CMS



<https://opendata.cern.ch>



<https://www.reana.io>