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

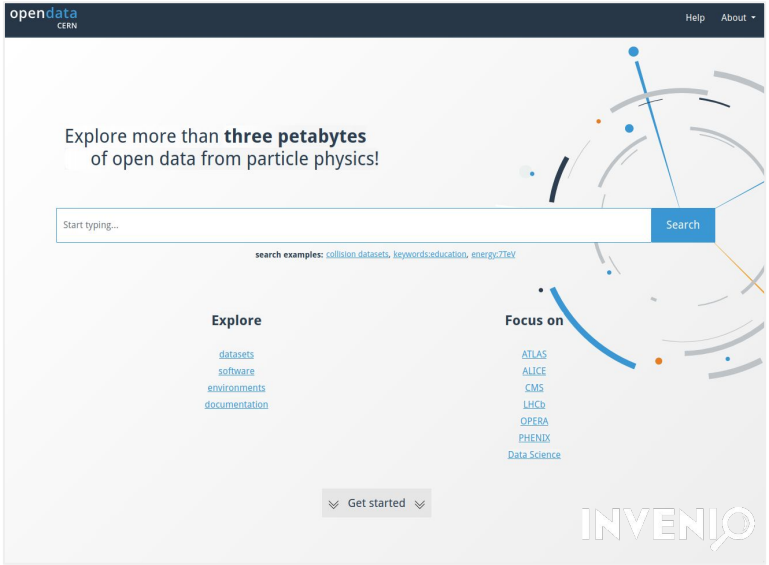
Digital repository for event-level particle physics open data

- 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)

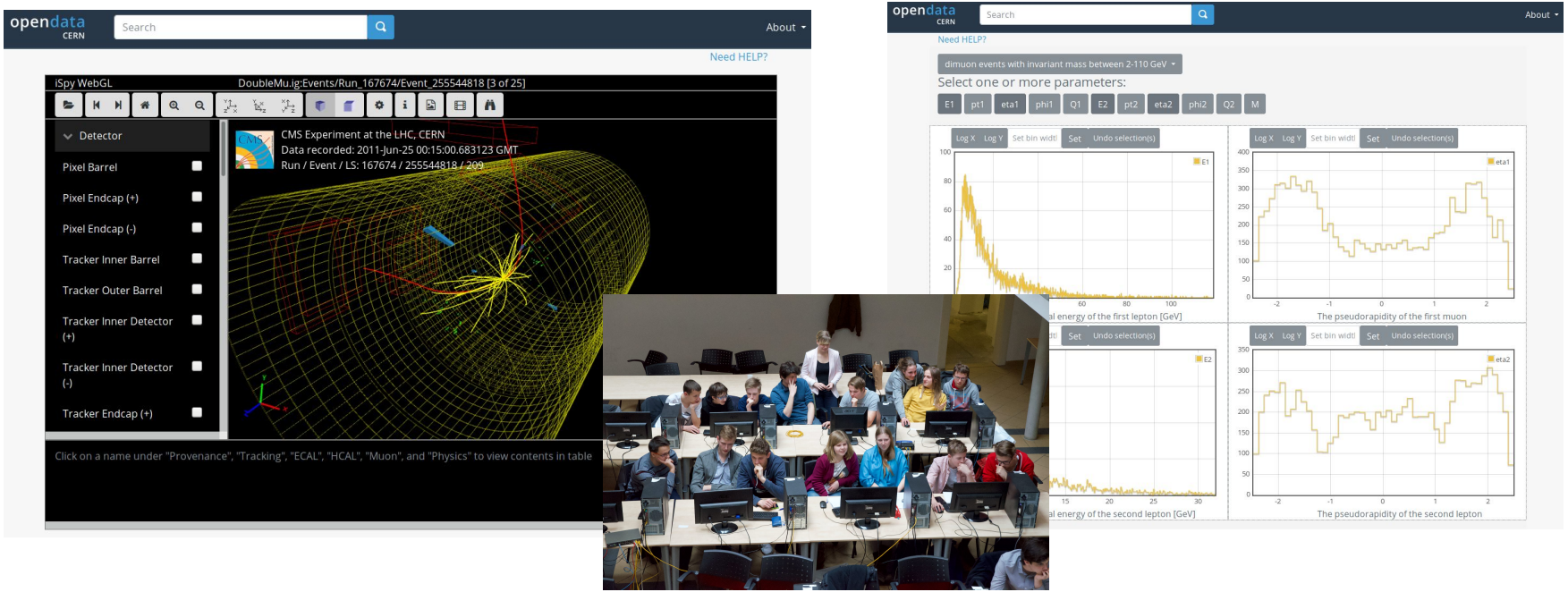
- over 15 000 bibliographic records
- over 1 500 000 files
- over 3 petabytes



https://opendata.cern.ch

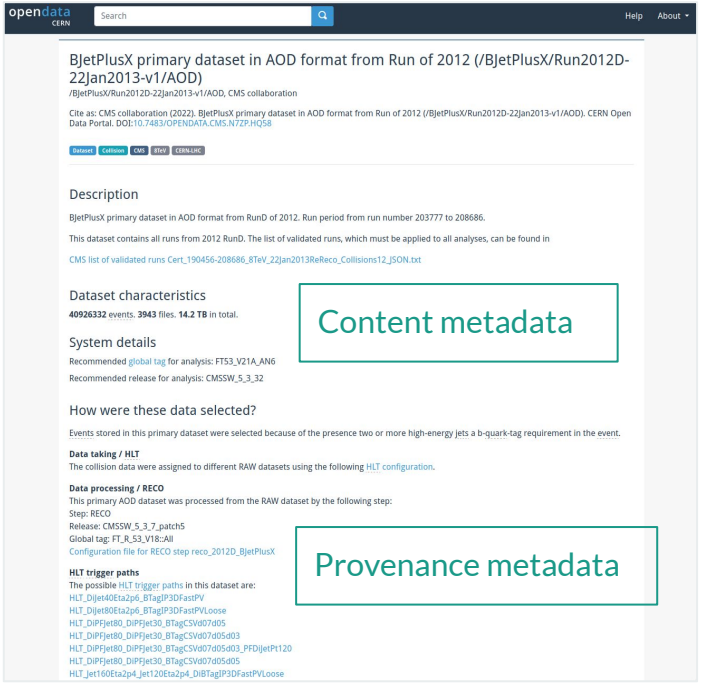
Developed by CERN in close collaboration with LHC (and non-LHC!) experiments

Education-oriented use cases



Interactive collision event displays and basic histogramming from derived datasets

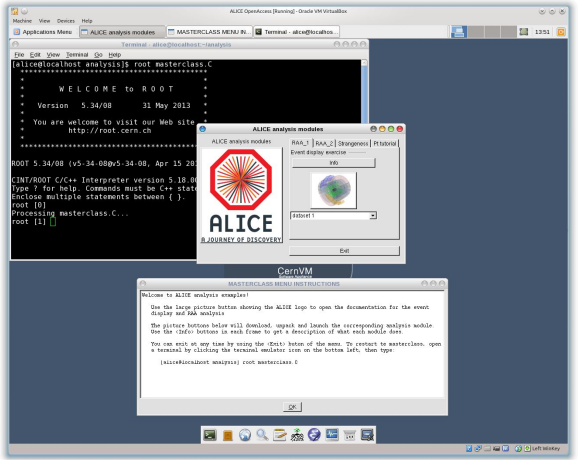
Research-oriented use cases



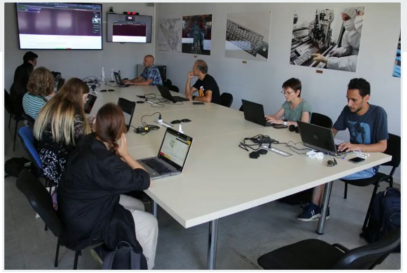
Content metadata

Provenance metadata

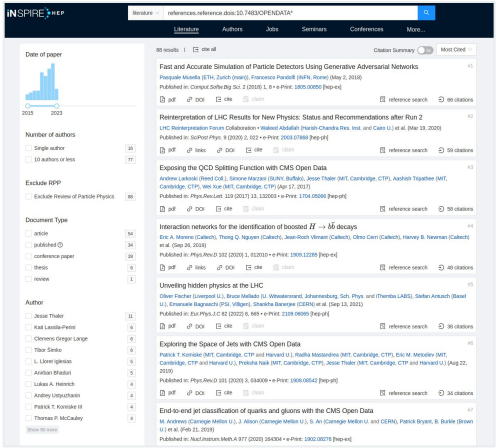
Explore research-grade primary datasets



Run virtual machines and containers with the same physics environment



CMS open data workshops for research use

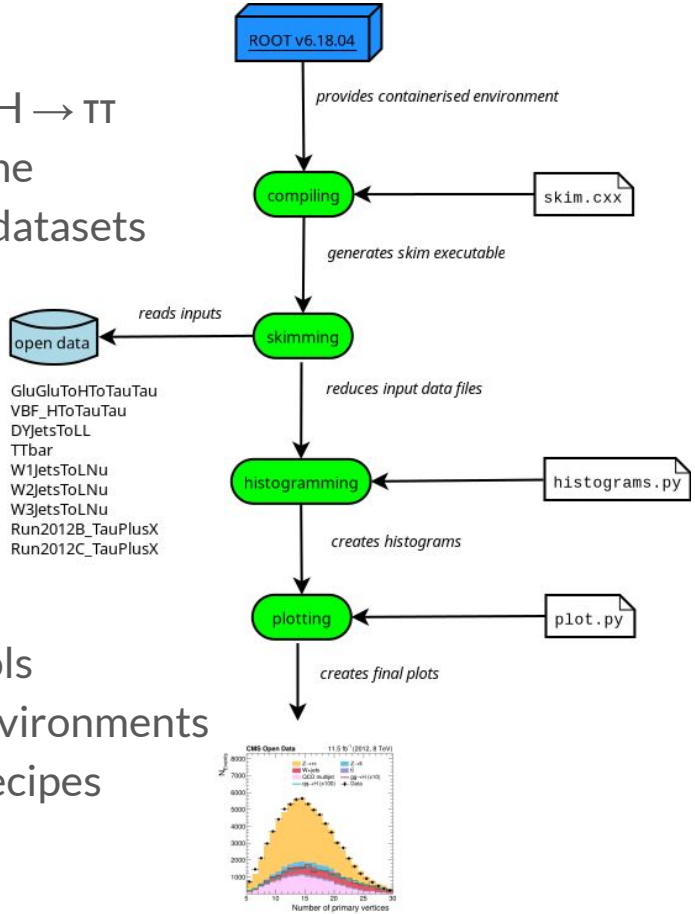


Enables independent research

How to use the data?

Data is accompanied with analysis examples

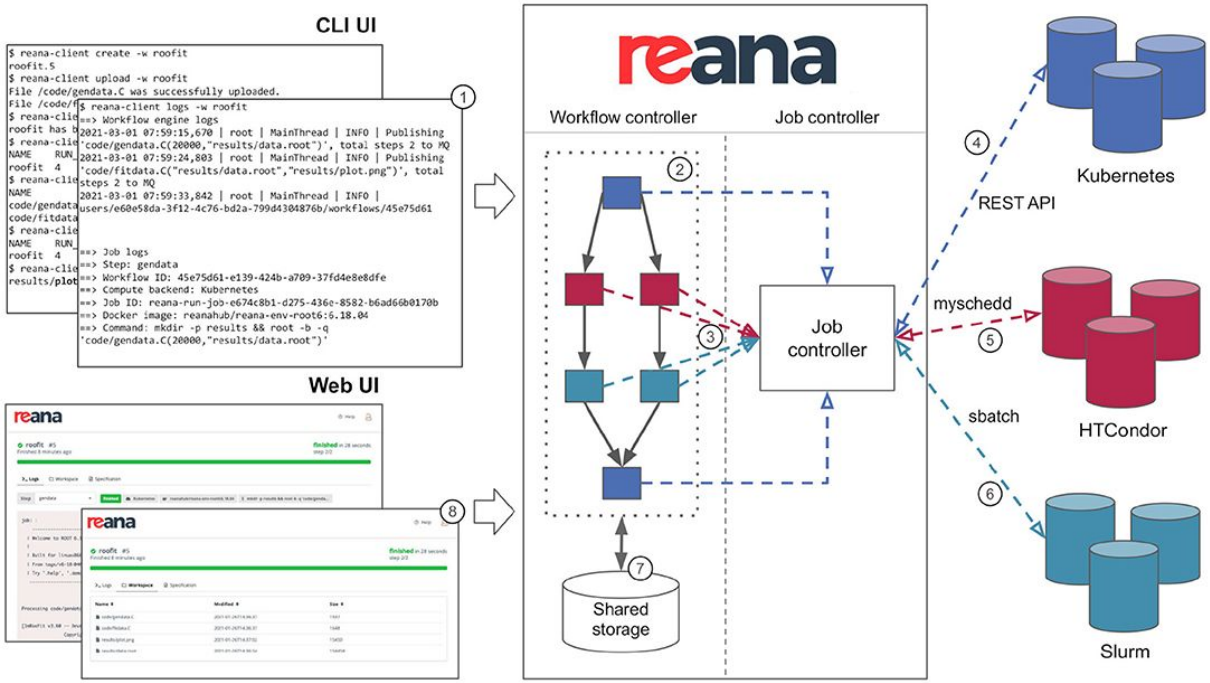
An example studying H → ττ lepton decays uses nine published CMS open datasets



Data is not enough!

- software and tools
- containerised environments
- computational recipes

REANA reusable analyses



Run containerised analysis workflows on the cloud

Multiple **compute backends**:

- Kubernetes
- HTCondor
- Slurm

Different **workflow languages**:

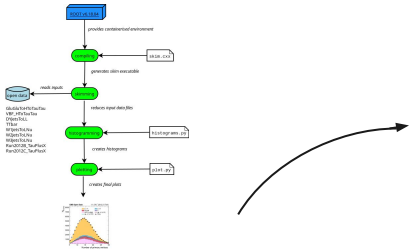
- CWL
- Serial
- Snakemake
- Yadage

Multiple **means of use**:

- Command-line client
- Web UI

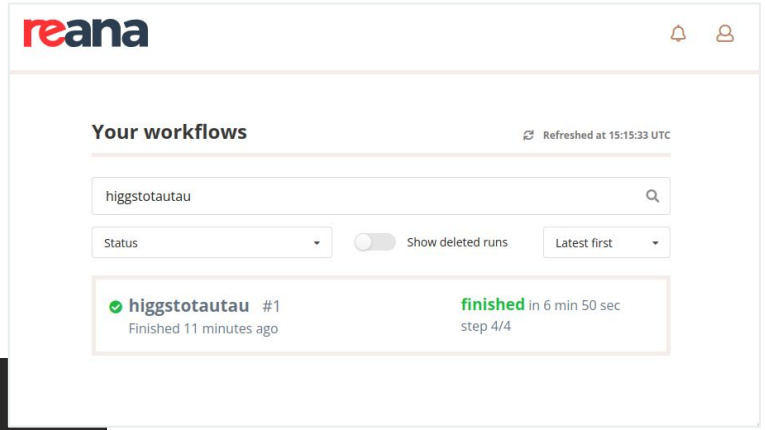
https://www.reana.io

H → ττ example running on REANA

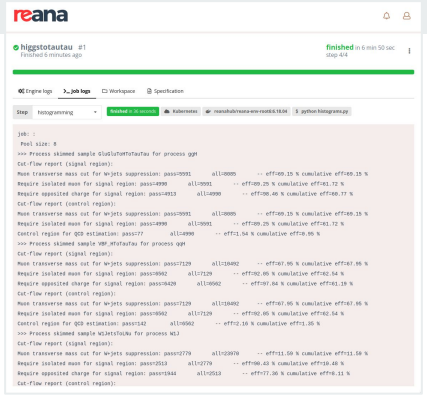


```
version: 0.6.0
inputs:
  files:
    - skim.cxx
    - histograms.py
    - plot.py
workflow:
  type: serial
  specification:
    steps:
      - name: compiling
        environment: reanahub/reana-env-root6:6.18.04
        commands:
          - g++ -g -O3 -Wall -Wextra -Wpedantic -o skim skim.cxx `root-config --cflags --libs`
      - name: skimming
        environment: reanahub/reana-env-root6:6.18.04
        commands:
          - ./skim
      - name: histogramming
        environment: reanahub/reana-env-root6:6.18.04
        commands:
          - python histograms.py
      - name: plotting
        environment: reanahub/reana-env-root6:6.18.04
        commands:
          - python plot.py
```

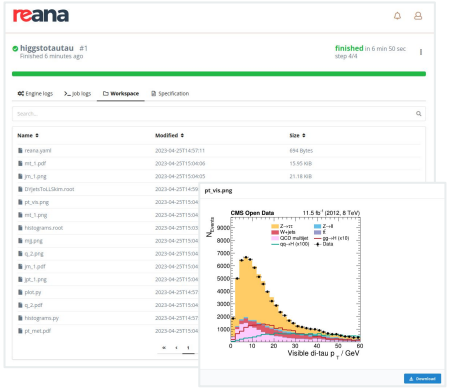
Structure data usage workflow



Run example on REANA



Consult workflow logs



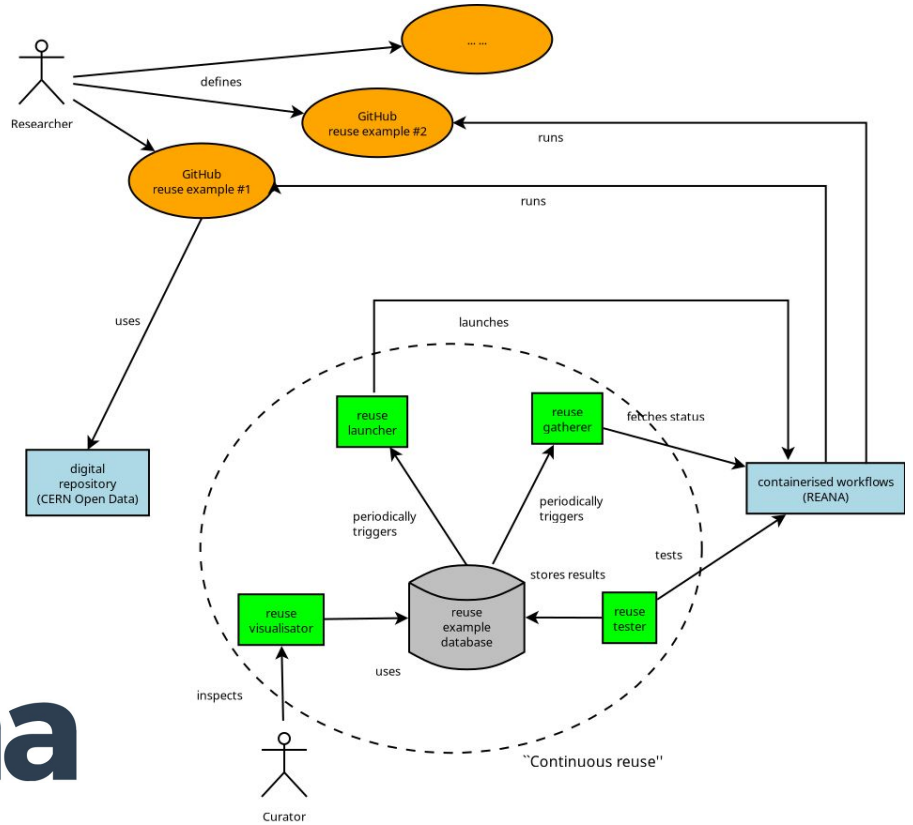
Visualise workflow outputs

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



opendata CERN

"continuous reuse"

reana

H → ττ 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

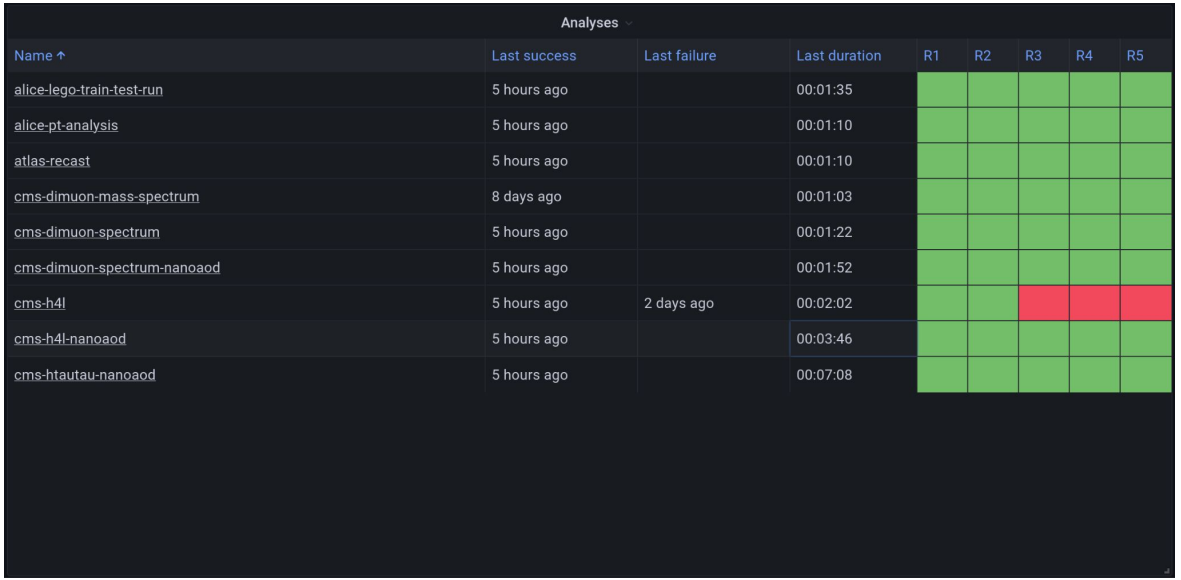
```
Feature: cms-htautau-nanoaod

  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 → τ τ 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

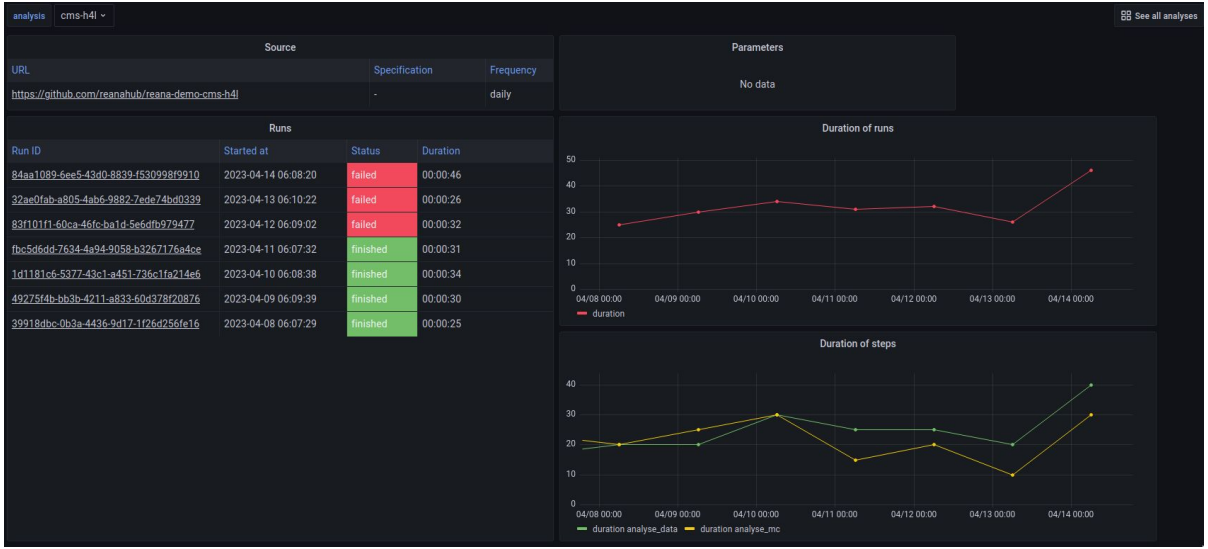


Continuous reuse dashboard home page showing the overall status of reuse analyses

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

opendata CERN

https://opendata.cern.ch

reana

https://www.reana.io