

CERN Open Data
CERN Analysis Preservation
REANA Reproducible Analyses

A brief status

Tibor Simko - CERN IT
Pamfilos Fokianos - CERN SIS

CERN Open Data

Status: *production* (since November 2014)

Size: 7K records, 800K files, 2 PB size

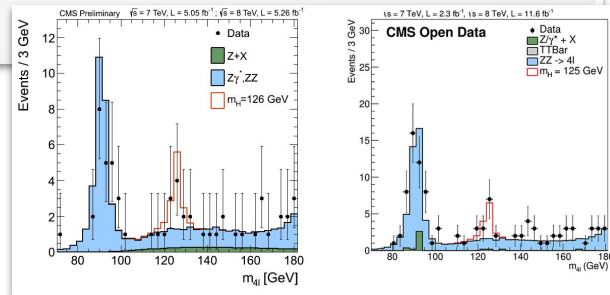
Purpose: “big data” sharing of event-level particle physics data and accompanying code for both education and research purposes

Content: raw samples, collision & simulated & derived datasets, docs, configs, software tools, example analyses, VMs, event display

Community: ALICE, ATLAS, CMS, LHCb, OPERA (coming: JADE, Data Science)

Notes: independent expert curation; batch ingestion workflows with Collaborations

The screenshot shows the CERN Open Data website. At the top left is the 'opendata CERN' logo, and at the top right is an 'About' link. The main heading reads 'Explore more than 1 petabyte of open data from particle physics!'. Below this is a search bar with the placeholder text 'Start typing...' and a 'Search' button. Underneath the search bar, it says 'search examples: collision datasets, keywords:education, energy:ZTeV'. There are two columns of links: 'Explore' with links for 'datasets', 'software', 'environments', and 'documentation'; and 'Focus on' with links for 'ATLAS', 'ALICE', 'CMS', and 'LHCb'. On the right side, there is a decorative graphic of a particle detector cross-section.



<http://opendata.cern.ch>

CERN Analysis Preservation

Status: *pilot*

Purpose: safe digital repository for preserving all individual user analysis assets of interest that are not stored anywhere (n-tuples, intermediate data, user code, plotting macros...)

Usage: describe analysis + deposit assets via CLI & Web UI + share with colleagues = preserve knowledge

Community: pilot examples with ALICE, ATLAS, CMS, LHCb; synergies with FREYA EC project partners

Notes: focus on “preservation” of analysis knowledge & assets without necessarily implying “reproducibility”; interconnection with Collaboration DBs and workflows

Resources: shared EOS space

The screenshot displays the CERN Analysis Preservation web application. At the top, there's a search bar and navigation icons. Below, a circular gauge shows '17 Total' items, with a breakdown of 'Your Drafts: 13' and 'Published: 4'. The 'DRAFTS' section lists several search entries with their titles and update times. The 'RECENTLY PUBLISHED IN COLLABORATION' section shows a list of search results. The 'WORKFLOWS' section displays a table with columns for workflow name, description, and status. A modal window is open, showing 'CADI STATUS' and 'CMS WG' filters.

```
$ cap-client files list --pid/p <existing pid>
[
  {
    "checksum": "md5:f0428126e7cf7b0d4af7091c68ae2a9f",
    "filename": "file.json",
    "filesize": 25,
    "id": "25852e50-be6d-47a5-897b-1f3df015fac7"
  },
  {
    "checksum": "md5:926fb9c44251d70614ee42d34c5365b6b",
    "filename": "Analysis_Notes_07112019.pdf",
    "filesize": 160898,
    "id": "89743c9b-106d-4235-8e96-23a164c7b1f4"
  }
]
```

<http://analysispreservation.cern.ch>

REANA Reproducible Analyses

Status: *pilot*

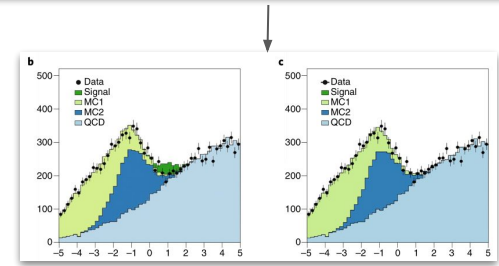
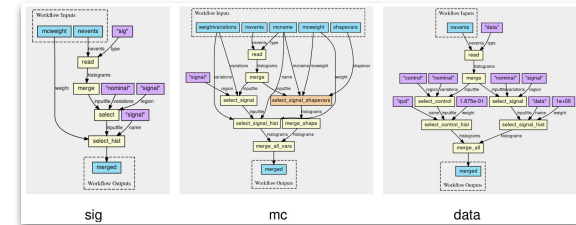
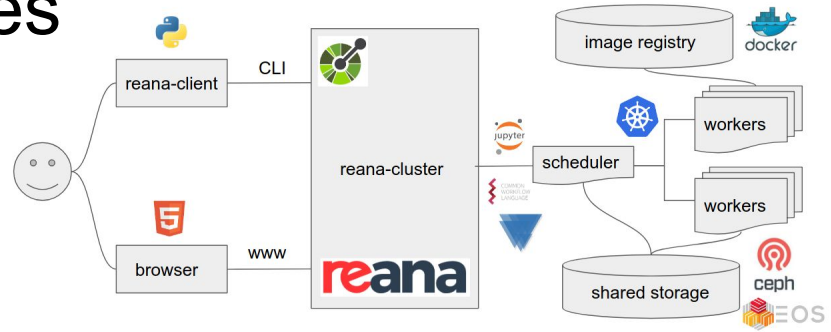
Purpose: run containerised scientific workflows on diverse compute clouds (Kubernetes, Condor, Slurm)

Usage: data + code + environment + workflow = reproducibility

Community: pilot examples with ALICE, ATLAS, CMS, FCC, LHCb; synergies with astronomy, life sciences, machine learning

Notes: focus on “preproducibility” of analysis during its active phase; structure analysis in a reproducible way to facilitate its future “preservation”

Resources: shared Ceph storage and Kubernetes cluster; need for proper experiment accounting



<http://www.reana.io>

Preservation & Reproducibility

