THE CERN ANALYSIS PRESERVATION PORTAL

IBERGRID 2019

Lara Lloret Iglesias
Instituto de Física de Cantabria

Overview

- Open Data at CERN → Why releasing data?
- Data Preservation Policies and goals of the different LHC experiments
- CERN Analysis Preservation Portal
- REANA

Why releasing data?

- **Inclusiveness:**

Science should be **open** and **inclusive**

- **Engagement:**

Being able to access data and *play* with it **engages** people with public research

- **Education**:

Teaching **students** about particle physics and data analysis in general

- **Society:**

A way to **give something back** to society (that funds our research ☺)

- **Scientific Research on Open Data**

Enhance scientific output by allowing non-collaboration members to access original data

How well are we doing right now?

- High Energy Physics is **doing well with *immediate* metadata**, such as: beam conditions, event and run numbers, provenance information (processing and reconstruction chain, software versions) **recorded together with data at time of data set creation**.
- **Doing worse on *context* metadata**, such as : how to pick up the right objects in the data and their documentation, how to know if there are additional selections, corrections…
- The information is readily available and even **obvious at time of immediate data** analysis, but then **easily forgotten**

Open Data and Analysis Preservation help/force us to meet this challenge

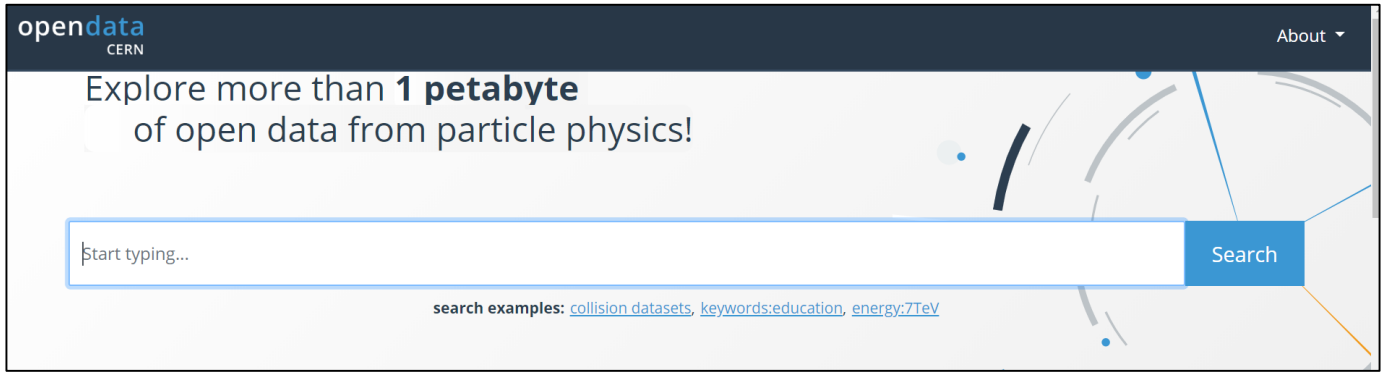
Information must be collected and released together with the data

CERN open data portal

opendata.cern.ch

- Access point to **growing range of data** produced through research at CERN.
- Disseminates preserved output from various research activities, including accompanying software and documentation needed to **understand and analyze the data** being shared.
- Adheres to established **global standards in data preservation** and Open Science

The products are shared under **open licenses** and issued with a digital object identifier (DOI) to make them citable.



- Close collaboration between experiments, CERN IT and scientific information services
- Many research and educational applications

Data access policy

ATLAS: http://opendata.cern.ch/record/413

LHCb: http://opendata.cern.ch/record/410 → **50% after 5 years**

ALICE: http://opendata.cern.ch/record/412 → **10% after 5 years**

"New" CMS policy data (20th of April 2018): 10.7483/OPENDATA.CMS.7347.JDWH

Release **up to 100% of collision** data within 10 years of data taking (in addition to the current practice of releasing 50% after 3 years)

CMS is the only experiment that has actually released research level data: this currently covers 50% of the Run I data, in line with the CMS policy.

Data access policy

Open Data levels defined in the Data Access Policies:

- **Level 1:** Public results → Papers, HepDATA, etc…
  - Working on providing capability for reinterpretation of searches
- **Level 2:** Outreach and Education and simplified data formats
  - **Lightweight environments** to allow the easy exploration of these data
  - **Full complexity may also be provided** for educational purposes.
- **Level 3:** Reconstructed Data → What is used by analysts
  - Embargo period allowing the collaboration members to perform the analyses
- **Level 4:** Raw Data → What the detector provides
  - Some samples could be released to the interest of data science (e.g. ML and AI)

The CERN analysis preservation portal (CAP)

- The CERN analysis preservation pursuits the following main goals:

➢ Describe and structure the **knowledge behind a physics analysis** aiming for its future reuse

  - Describe all the assets of an analysis and track data provenance.
  - Ensure sufficient documentation and capture associated links.

➢ Store information about the analysis **input data**, the **analysis code** and its dependencies, the **runtime computational environment** and the **analysis workflow** steps, and any other necessary dependencies in a trusted digital repository.

➢ Create a web portal with a **friendly structure** to both register and search the information

The CERN analysis preservation portal (CAP)

CAP is built on the **Invenio digital library** framework

(used in INSPIRE, CERN Open Data, Zenodo and many others)

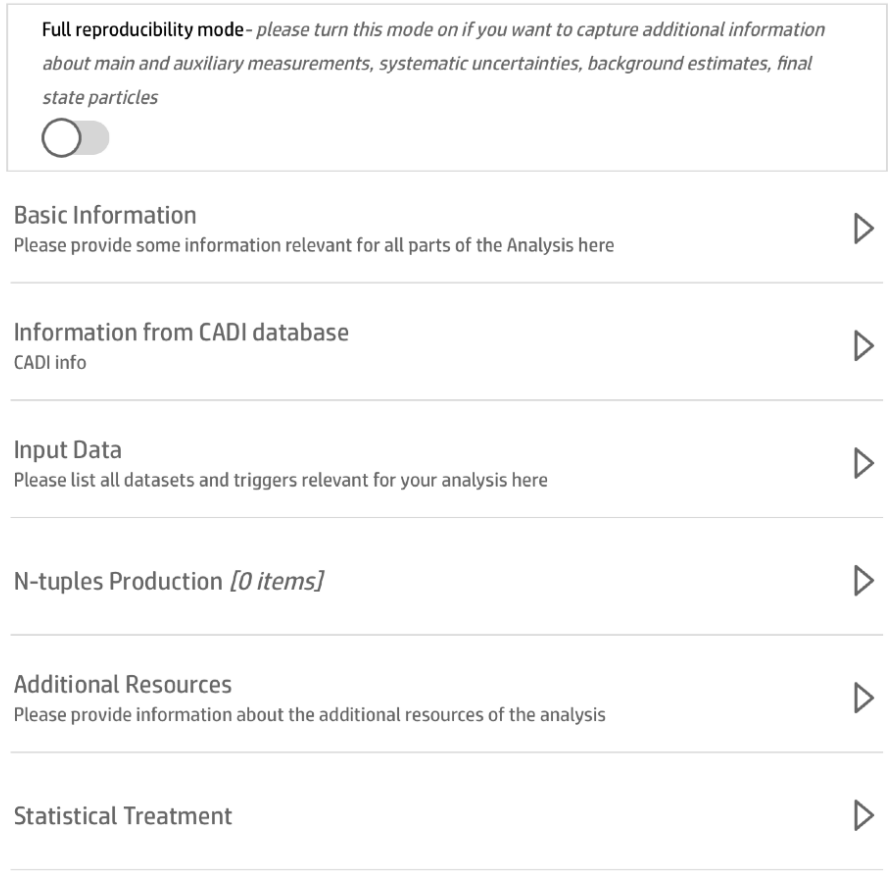
Data are modelled in **JSON format**

JSON Schema with standard metadata requirements, informed by existing databases and practices in the collaboration

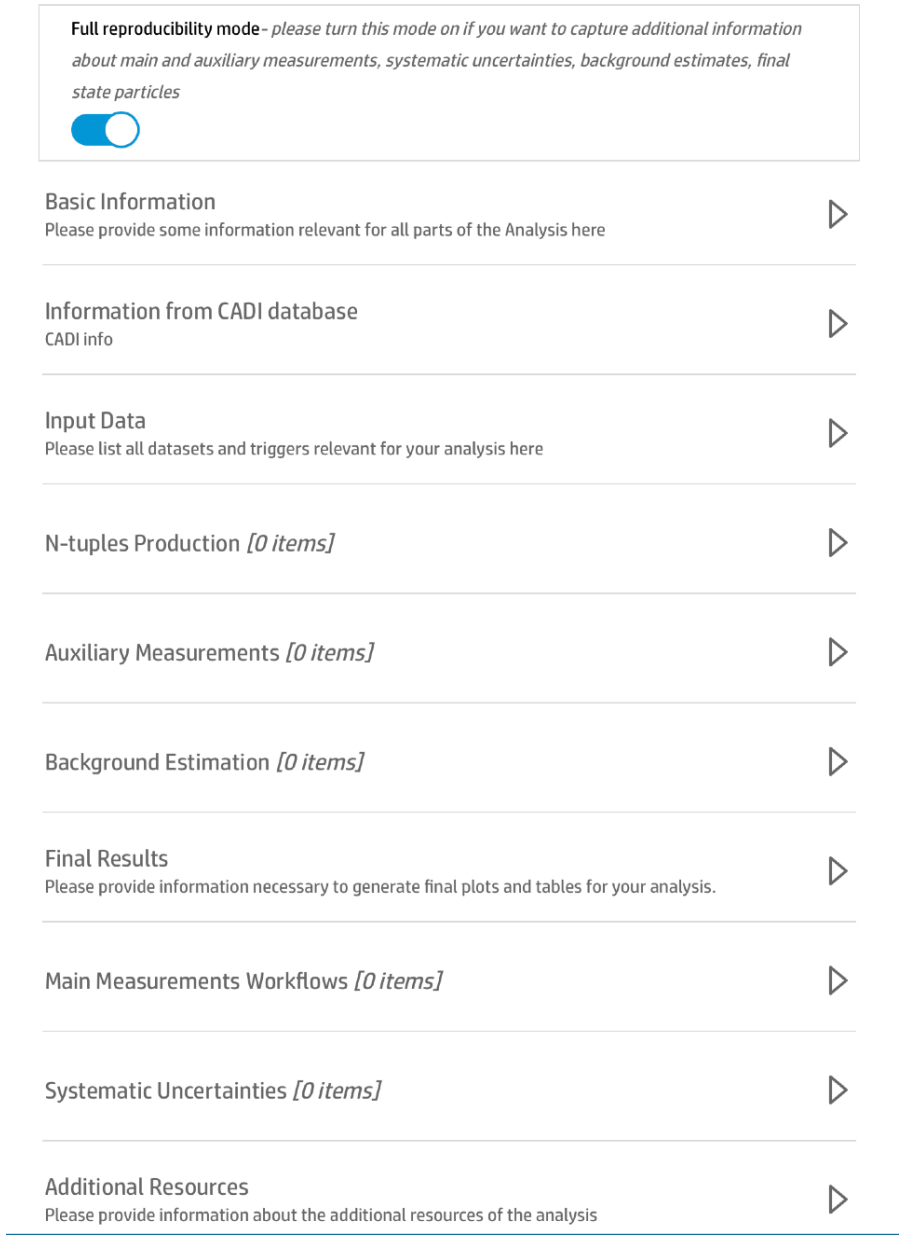
Elasticsearch cluster for indexing and information retrieval needs

Open Archival Information System (OAIS) practices to ensure **long-term preservation**

Vanilla Mode



Re-Use Mode



Two different analysis preservation modes:

- **Vanilla mode:** Basic assets preservation
- **Reuse mode:** Great level of detail → Aiming for reuse

SOME LINKS

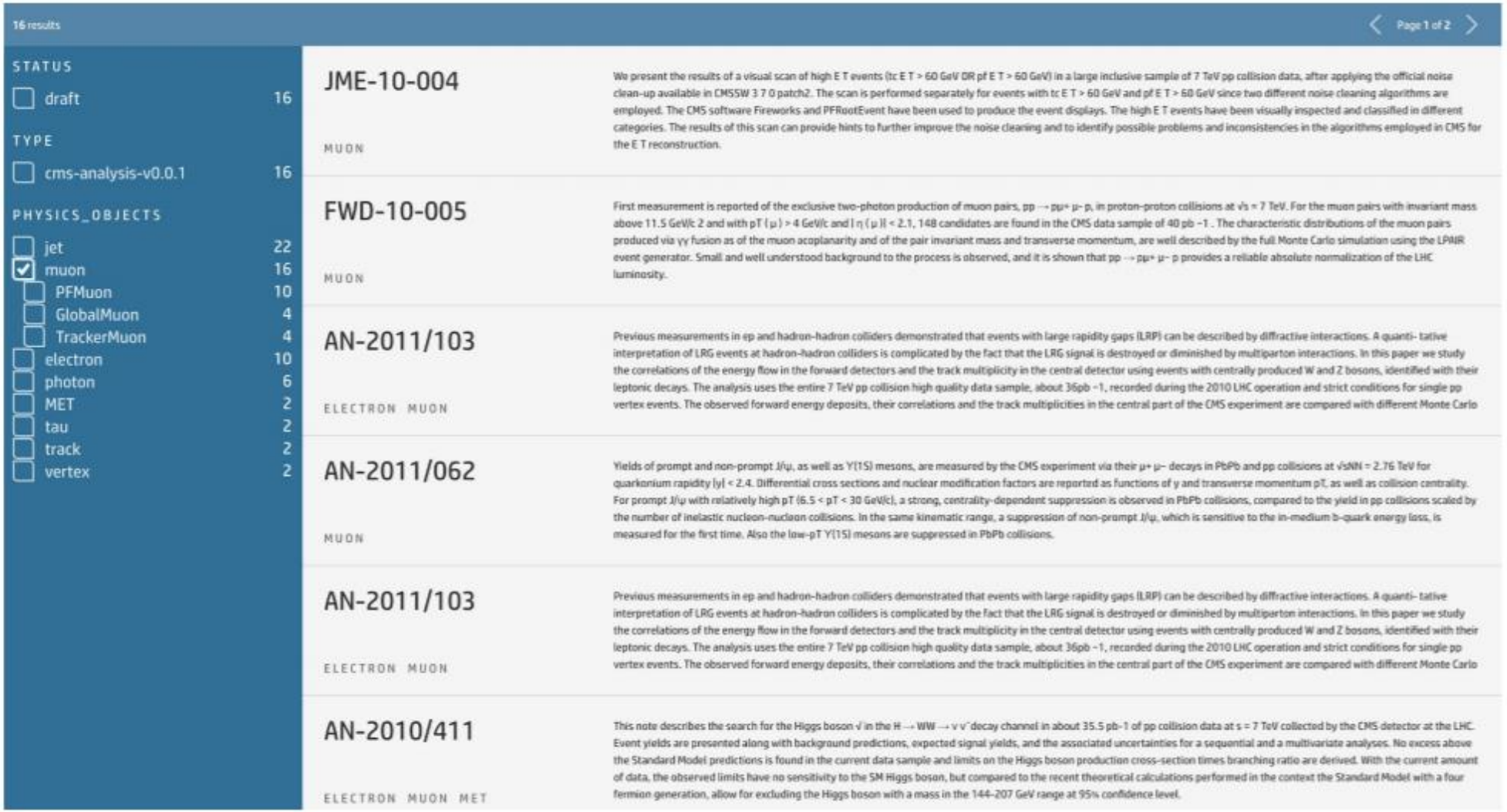
Access to the CERN Analysis Preservation service here:
https://analysispreservation.cern.ch

Intructions about how to submit n-tuples and output macros here:
https://cernanalysispreservation.readthedocs.io/en/latest/tutorials.html

Submission and retrieval of analysis material can be automatised via the command-line client:
https://cap-client.readthedocs.io/

People are already testing it with very positive feedback

Indexed - searchable



Command line client

Command line client to facilitate workflow automation:

- Allows to create/edit JSON files with the analysis information and submit it directly from the terminal
- Information propagated to the web portal

```
$ pip install cap-client
```

```
$ export CAP_SERVER_URL=https://analysispreservation.cern.ch/
$ export CAP_ACCESS_TOKEN=<your generated access token from server>
```

```
$ cap-client files upload <file path> --pid/-p <existing pid>

$ cap-client files upload file.json -p 89b593c498874ec8bcafc88944c458a7

File uploaded successfully.
```

REANA

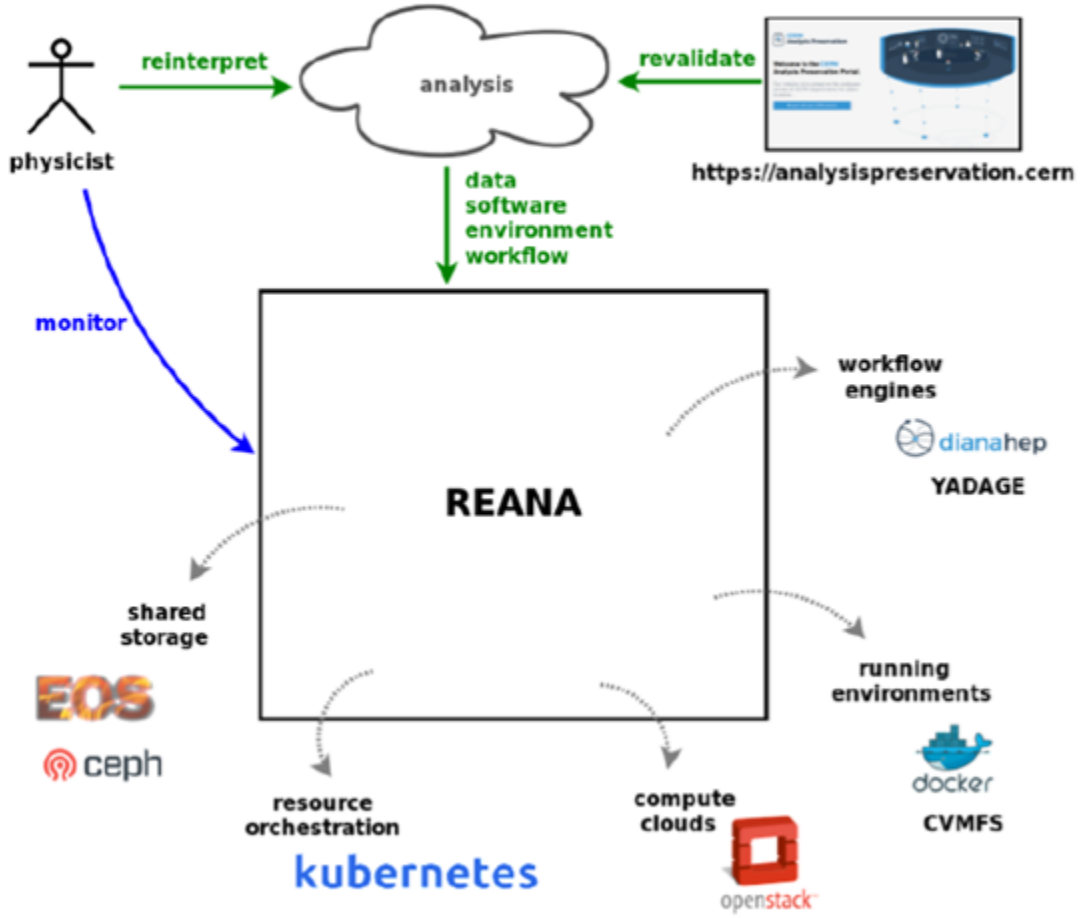
Reproducible research data analysis platform

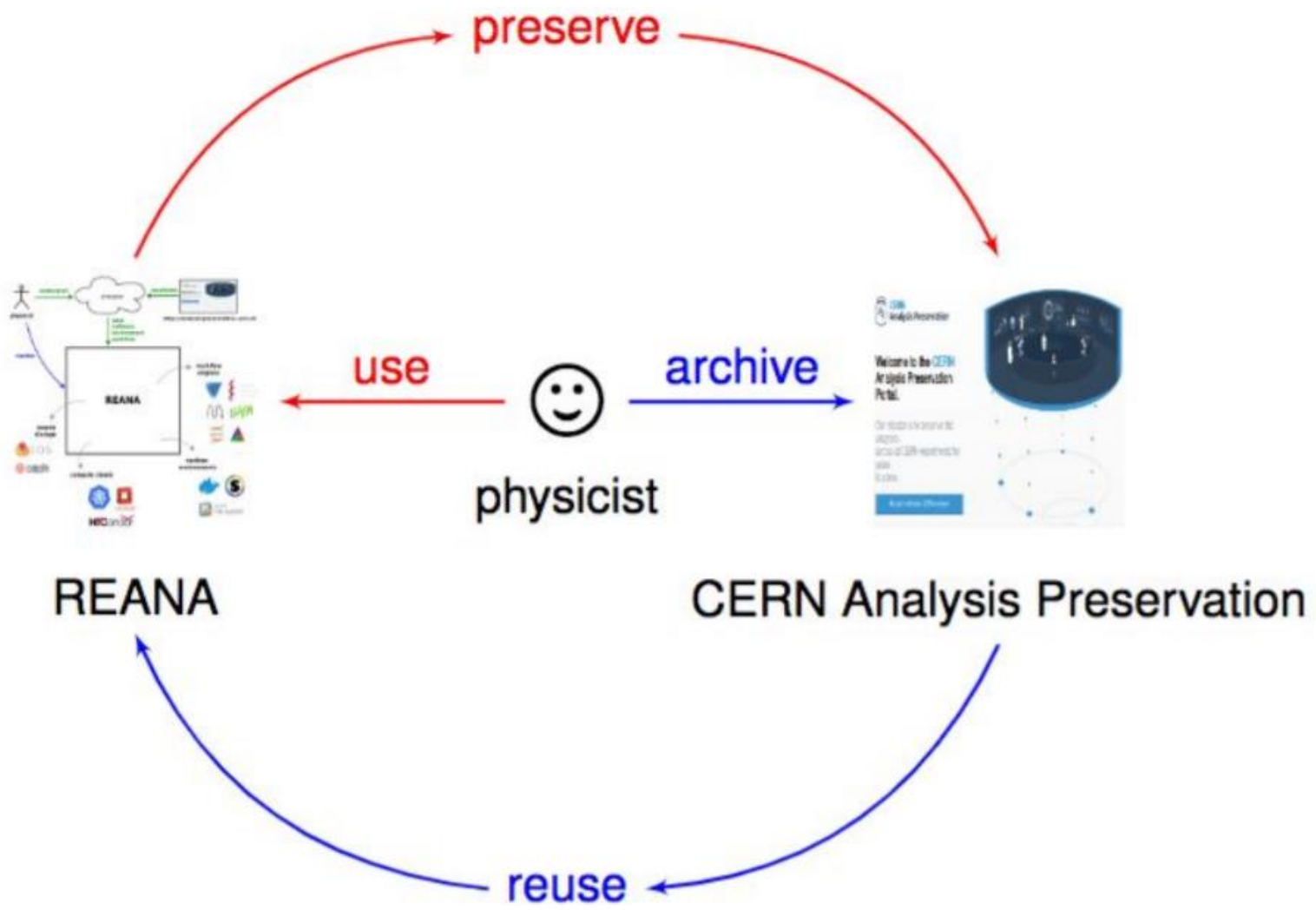
a system for **reusable analysis** execution **on the cloud**

https://reanahub.io

supporting **multiple scenarios**

- multiple computing clouds
  → CERN OpenStack
- multiple running environments
  → Docker with CVMFS
- multiple resource orchestration
  → Kubernetes
- multiple workflow engines
  → Yadage
- multiple shared storage systems
  → Ceph, EOS




preserve
use
archive
physicist
REANA
CERN Analysis Preservation
reuse

Conclusions

- **All four experiments are doing as much as they can to make the Open Data goals advance** → Limited by experiment resources
- Open Data enables us to **engage with the outside world** on a more meaningful level enabling us to **show how we work**
- There are plenty of **tools and datasets ready for you to use**
- The data can be used to **develop lab courses, full visualisations, create teaching materials**…and also **research results**.
- Working on **analysis preservation** and **reproducibility**
  - CERN Analysis Preservation Portal (CAP) already being filled
- REANA: Platform for reproducible research analysis based on dockers is in place
  - Next steps: Integration with CAP

FURTHER READING

nature physics

Perspective | Open Access | Published: 15 November 2018

Open is not enough

Xiaoli Chen, Sünje Dallmeier-Tiessen ✉, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, Diego Rodriguez Rodriguez, Tibor Šimko ✉, Tim Smith, Ana Trisovic ✉, Anna Trzcinska, Ioannis Tsanaktsidis, Markus Zimmermann, Kyle Cranmer, Lukas Heinrich, Gordon Watts, Michael Hildreth, Lara Lloret Iglesias, Kati Lassila-Perini & Sebastian Neubert

Nature Physics **15**, 113–119 (2019) | Download Citation ↓

Backup

The other experiments

ALICE APPROACH:

- ALICE LEGO trains: Centralized analysis system within ALICE to run analyses on the Grid
- Only 5 scripts → easy to automatize and preserve

ATLAS APPROACH:

- ATLAS has a already had quite extensive analysis tracking tool: "Analysis Glance" — recently expanded quiet a bit
- Used to track analysis through the entire lifecycle
- Large amount of Metadata
- Envisioned to be main input source to CAP → Need work to build ingestion route Glance → CAP

LHCb APPROACH:

- Two approaches:
  - Partial Reproducibility: Store ntuples and anything happening afterwards
  - Total Reproducibility: Needs full workflow and adecuate framework (docker)
- Different pipeline tools under evaluation