# Open Data at CERN

## UNIGE Open Science Course: Open Science at CERN

**Pablo Saiz**

**22 Nov 2023**

Slides taken from Tibor Šimko

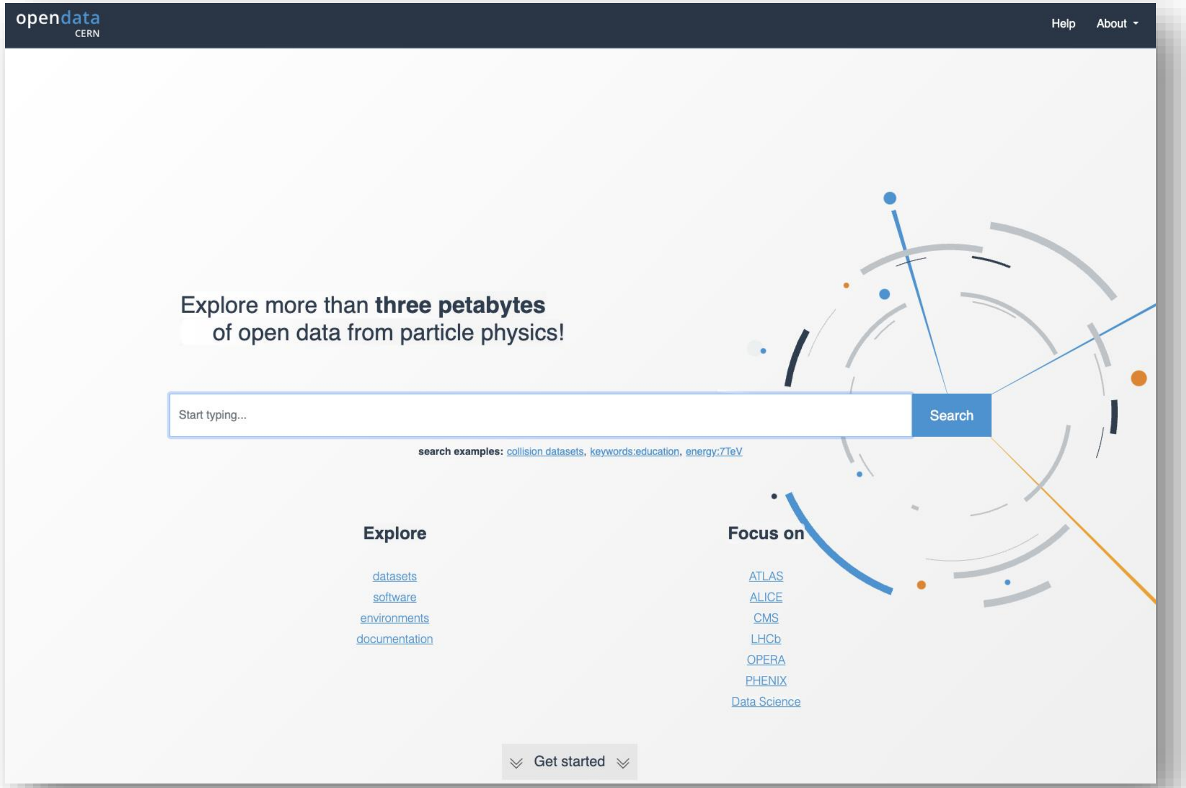
# Content

- CERN Open Data portal
- FAIR principle
- REANA
- Summary

# CERN Open Data (COD) Portal

http://opendata.cern.ch

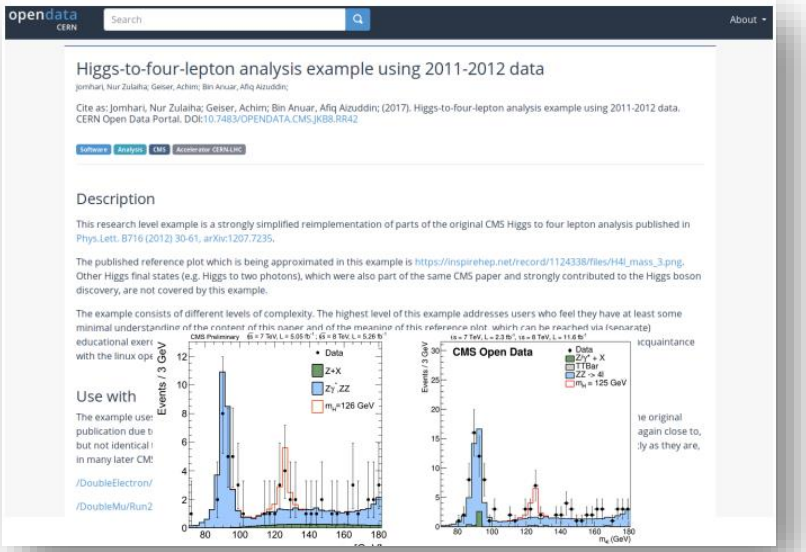
- **Repository of data**
  - Launched in November 2014
- **Plenty of content**
  - Dataset
    - Collision, simulated and derived
  - Documentation
    - Glossary, tutorials, configuration, examples
  - Software
    - Frameworks, virtual machines, containers
- **Current size (Nov 2023)**
  - > 17.000 records
  - > 1.900.000 files
  - > 4.5 PB



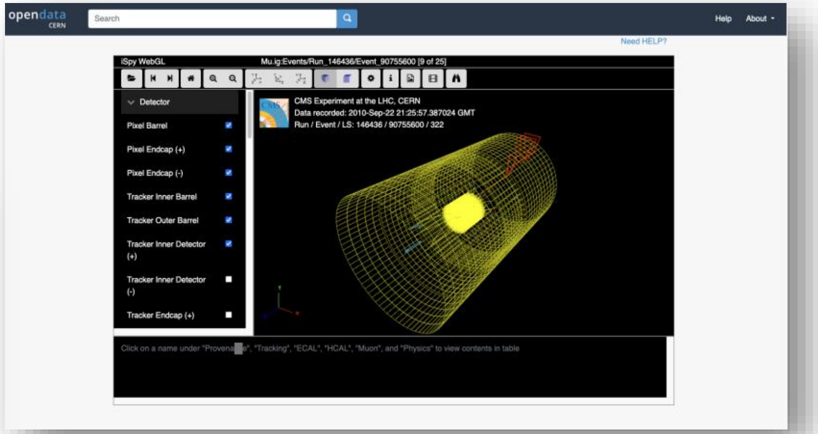

Developed by CERN in collaboration with Experiments

ALICE ATLAS CMS LHCb OPERA PHENIX

# COD Portal

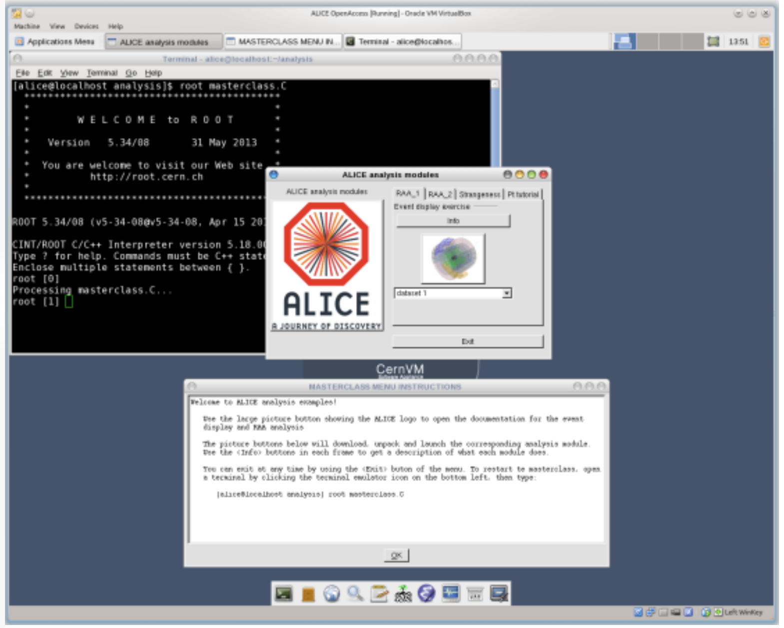



Run research-grade analysis examples



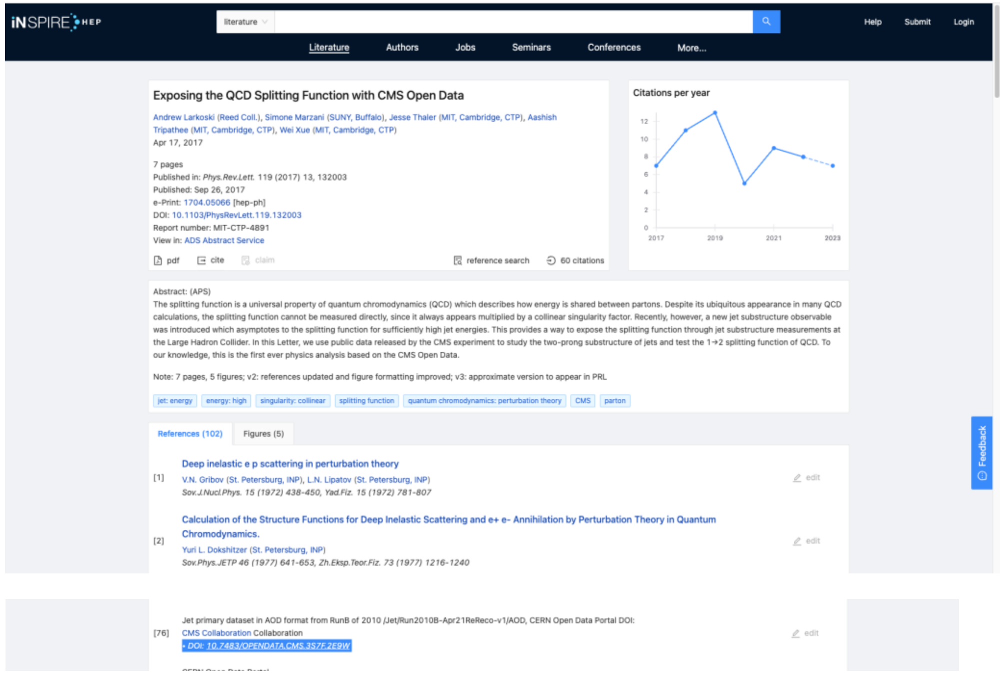


Interactive event display and histograms

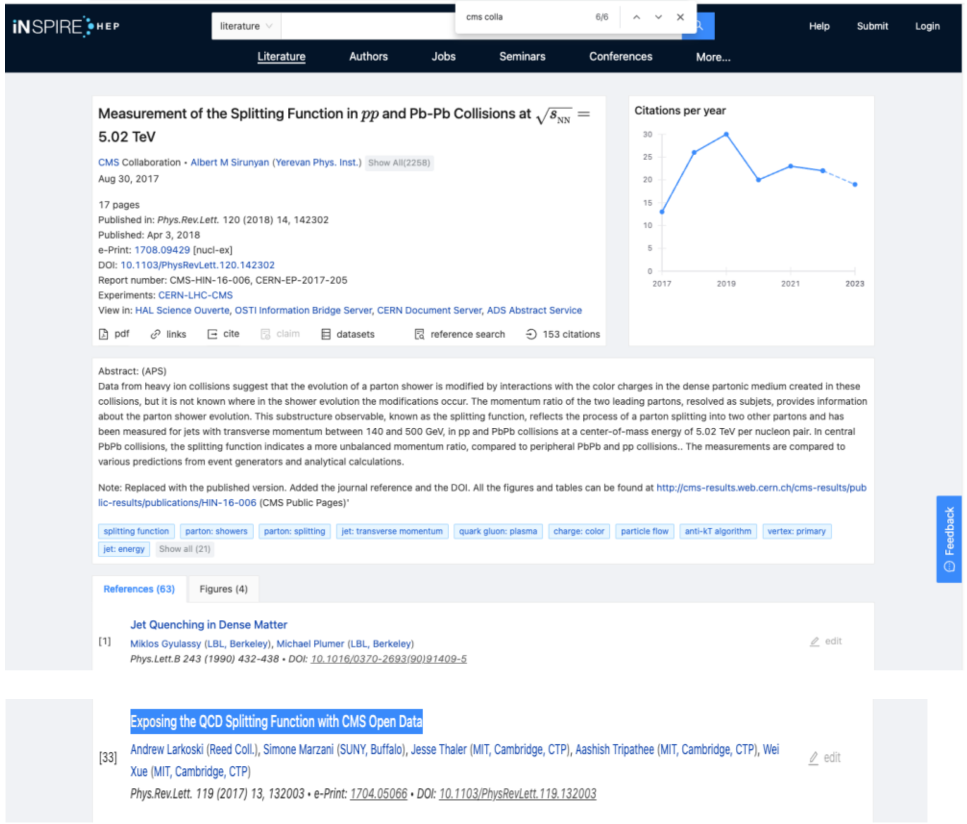



Run CernVM Virtual Machines

# Enables independent theoretical research

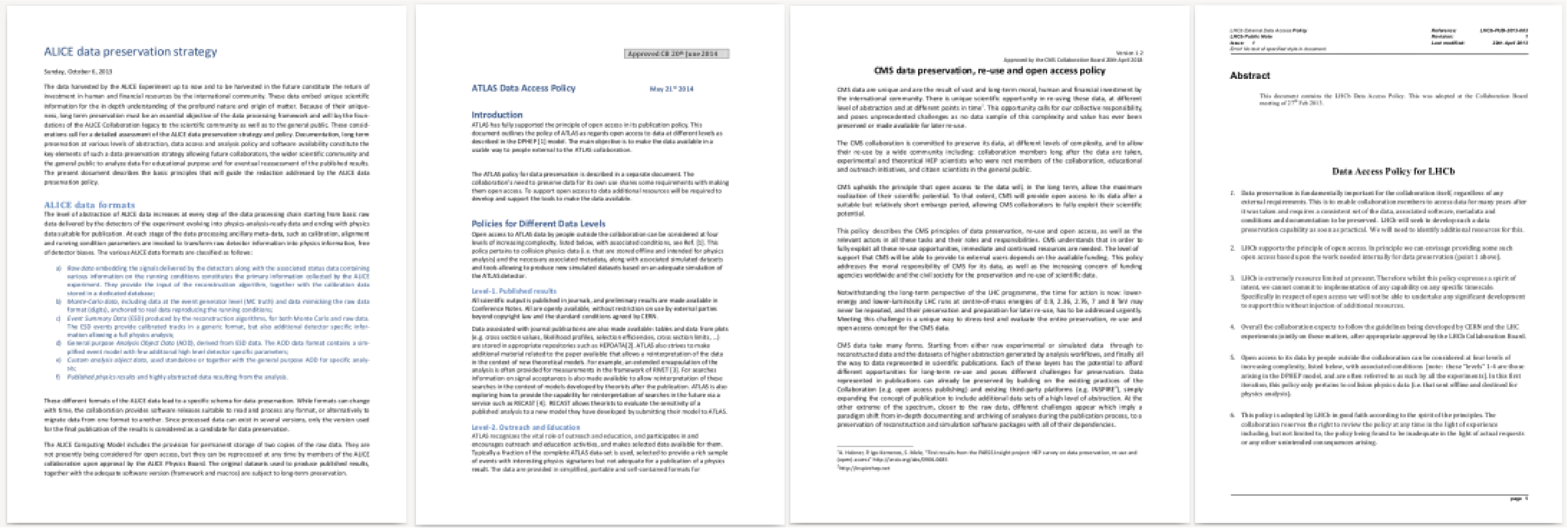



Over thirty papers citing CMS Open Data




… that the CMS collaboration cites

# LHC collaboration data preservation and open access policies


ALICE data preservation strategy

Approved CB 20th June 2014

ATLAS Data Access Policy

Introduction

Policies for Different Data Levels

CMS data preservation, re-use and open access policy

Abstract

Data Access Policy for LHCb

ALICE ATLAS EXPERIMENT CMS LHCb

**Restricted data → Embargo period (~5 years) → Open data**

# FAIR guiding principles

- **F**indable
- **A**ccesible
- **I**nteroperable
- **R**eusable

https://www.nature.com/articles/sdata201618

Open access | Published: 15 March 2016

## The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, ... Barend Mons + Show authors

*Scientific Data* **3**, Article number: 160018 (2016) | Cite this article



An Addendum to this article was published on 19 March 2019

### Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

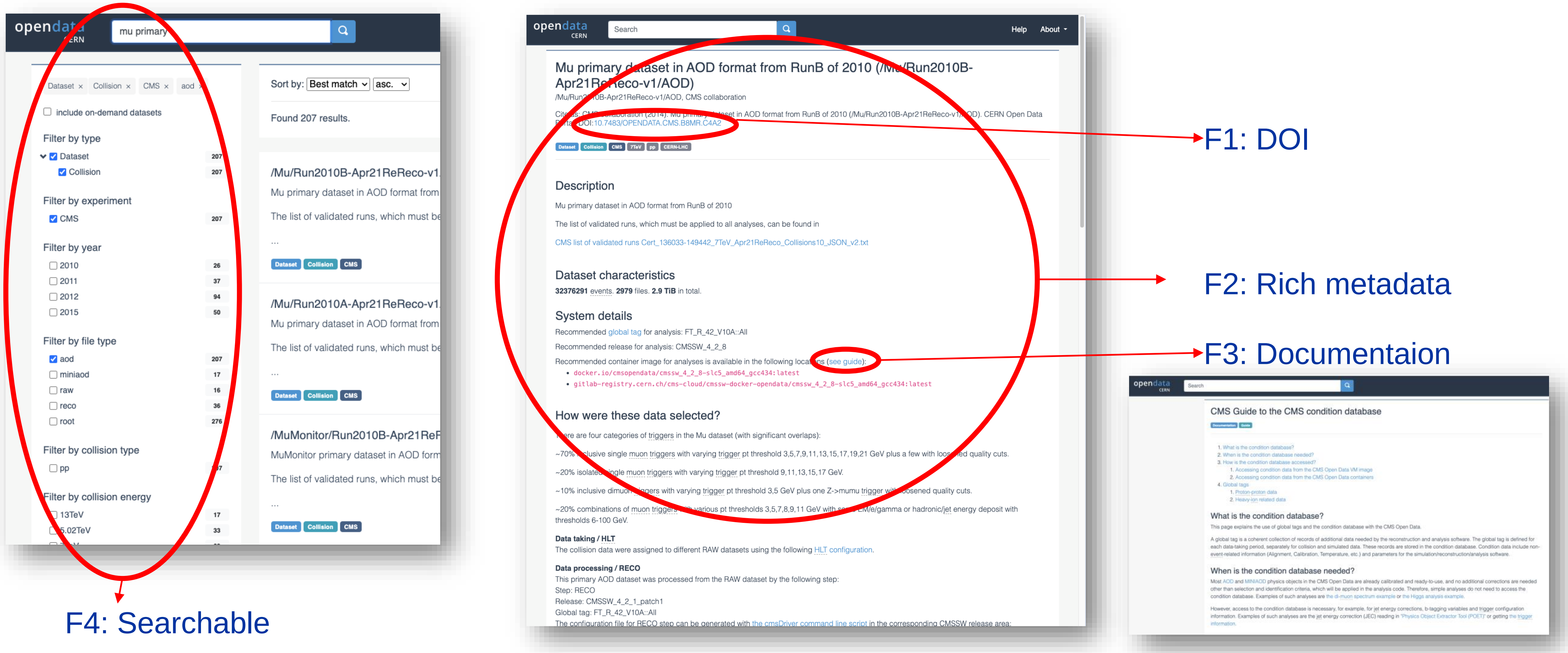
# FAIR: Findable

Data and metadata should be easy to find by humans and computers

Principles:

- F1: (meta)data are assigned a globally unique and persistent identifier
- F2: data are described with rich metadata (defined by R1 below)
- F3: metadata clearly and explicitly include the identifier of the data it describes
- F4: (meta)data are registered or indexed in a searchable resource

# COD Findable



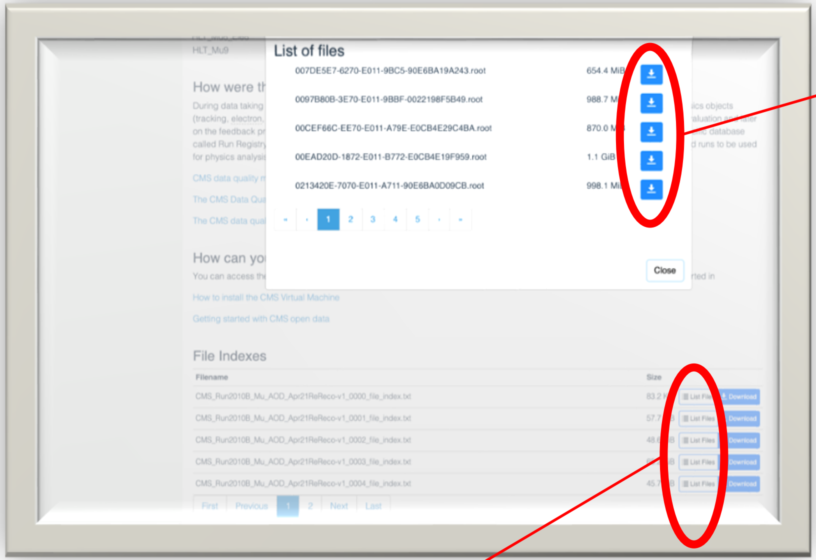
# FAIR: Accessible

Once the users find the required data, they need to know how it can be access

Principles:

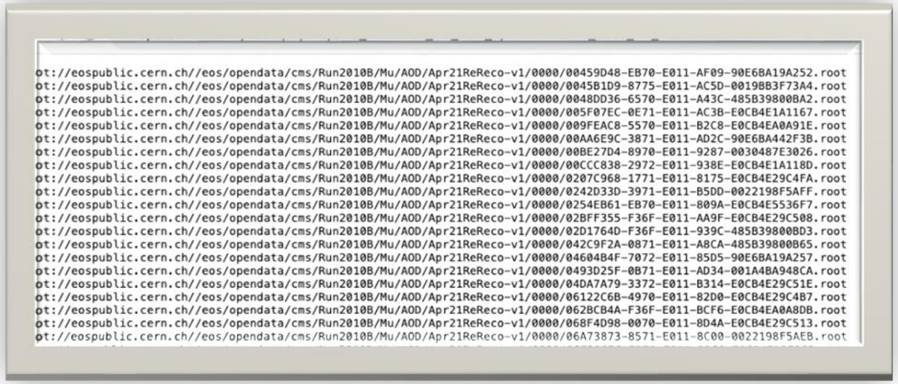
- A1: (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

# COD: Accessible



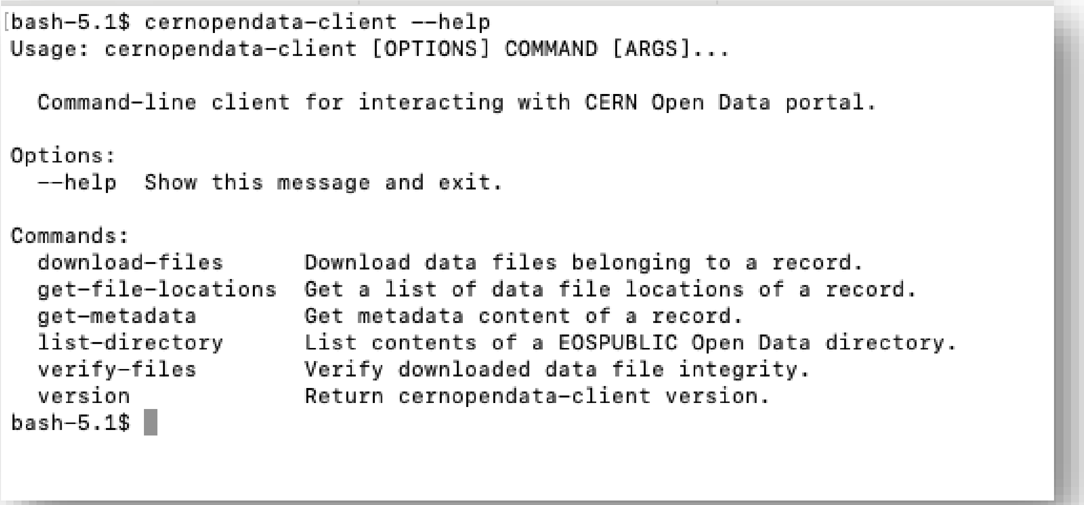

A1: Retriavable

Big datasets



```
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00459D48-EB70-E011-AF09-90E6BA19A252.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0045B1D9-8775-E011-AC5D-0019BB3F73A4.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0048DD36-6570-E011-A43C-485B39800BA2.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/005F07EC-0E71-E011-AC3B-E0CB4E1A1167.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/009FEAC8-5570-E011-B2C8-E0CB4EA0A91E.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00AA6E9C-3871-E011-AD2C-90E6BA442F3B.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00BE27D4-8970-E011-9287-0030487E3026.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00CCC838-2972-E011-938E-E0CB4E1A118D.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0207C968-1771-E011-8175-E0CB4E29C4FA.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0242D33D-3971-E011-B5DD-0022198F5AFF.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0254EB61-EB70-E011-809A-E0CB4E5536F7.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/02BFF355-F36F-E011-AA9F-E0CB4E29C508.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/02D1764D-F36F-E011-939C-485B39800BD3.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/042C9F2A-0871-E011-A8CA-485B39800B65.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/04604B4F-7072-E011-85D5-90E6BA19A257.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0493D25F-0B71-E011-AD34-001A4BA948CA.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/04DA7A79-3372-E011-B314-E0CB4E29C51E.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/06122C6B-4970-E011-82D0-E0CB4E29C4B7.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/062BCB4A-F36F-E011-BCF6-E0CB4EA0A8DB.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/068F4D98-0070-E011-8D4A-E0CB4E29C513.root
ot://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/06A73873-8571-E011-8C00-0022198F5AEB.root
```

https://github.com/cernopendata/cernopendata-client

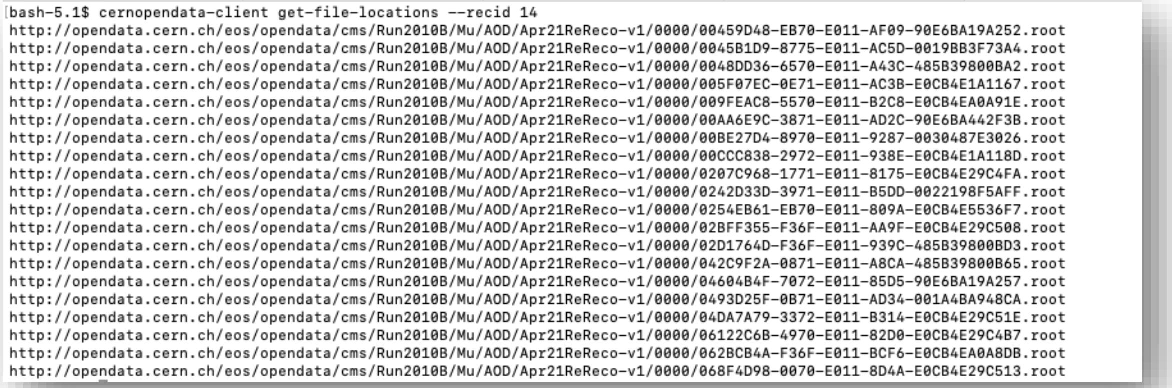


```
bash-5.1$ cernopendata-client --help
Usage: cernopendata-client [OPTIONS] COMMAND [ARGS]...

  Command-line client for interacting with CERN Open Data portal.

Options:
  --help  Show this message and exit.

Commands:
  download-files      Download data files belonging to a record.
  get-file-locations  Get a list of data file locations of a record.
  get-metadata        Get metadata content of a record.
  list-directory      List contents of a EOSPUBLIC Open Data directory.
  verify-files        Verify downloaded data file integrity.
  version             Return cernopendata-client version.
bash-5.1$
```



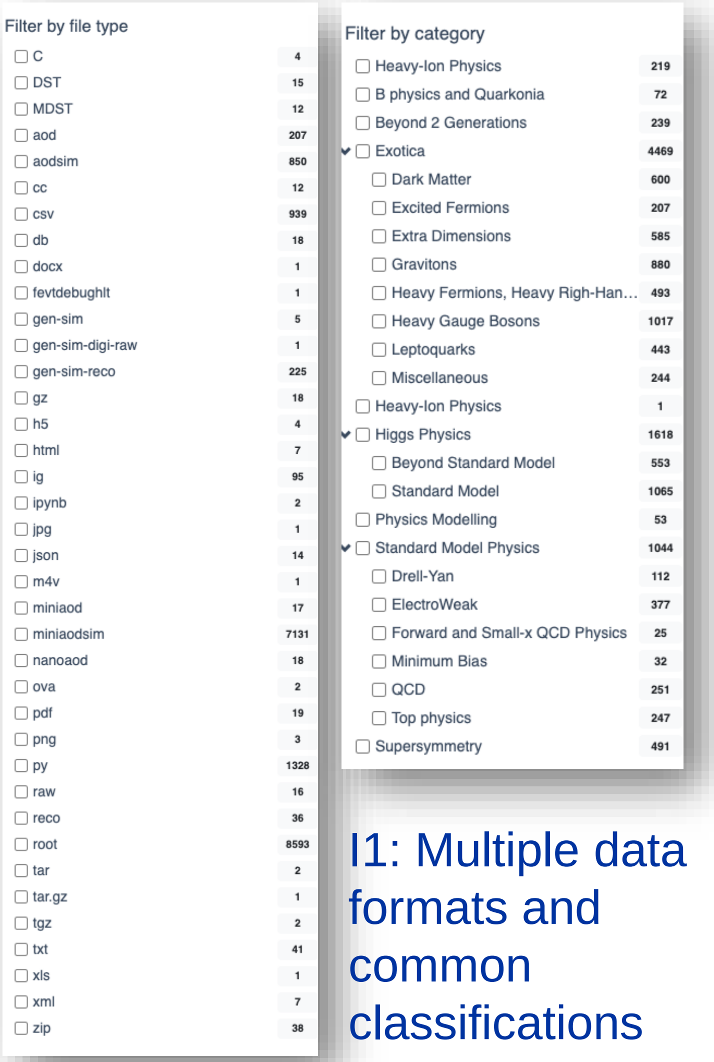
```
bash-5.1$ cernopendata-client get-file-locations --recid 14
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00459D48-EB70-E011-AF09-90E6BA19A252.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0045B1D9-8775-E011-AC5D-0019BB3F73A4.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0048DD36-6570-E011-A43C-485B39800BA2.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/005F07EC-0E71-E011-AC3B-E0CB4E1A1167.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/009FEAC8-5570-E011-B2C8-E0CB4EA0A91E.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00AA6E9C-3871-E011-AD2C-90E6BA442F3B.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00BE27D4-8970-E011-9287-0030487E3026.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00CCC838-2972-E011-938E-E0CB4E1A118D.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0207C968-1771-E011-8175-E0CB4E29C4FA.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0242D33D-3971-E011-B5DD-0022198F5AFF.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0254EB61-EB70-E011-809A-E0CB4E5536F7.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/02BFF355-F36F-E011-AA9F-E0CB4E29C508.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/02D1764D-F36F-E011-939C-485B39800BD3.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/042C9F2A-0871-E011-A8CA-485B39800B65.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/04604B4F-7072-E011-85D5-90E6BA19A257.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0493D25F-0B71-E011-AD34-001A4BA948CA.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/04DA7A79-3372-E011-B314-E0CB4E29C51E.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/06122C6B-4970-E011-82D0-E0CB4E29C4B7.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/062BCB4A-F36F-E011-BCF6-E0CB4EA0A8DB.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/068F4D98-0070-E011-8D4A-E0CB4E29C513.root
```

# FAIR: Interoperable

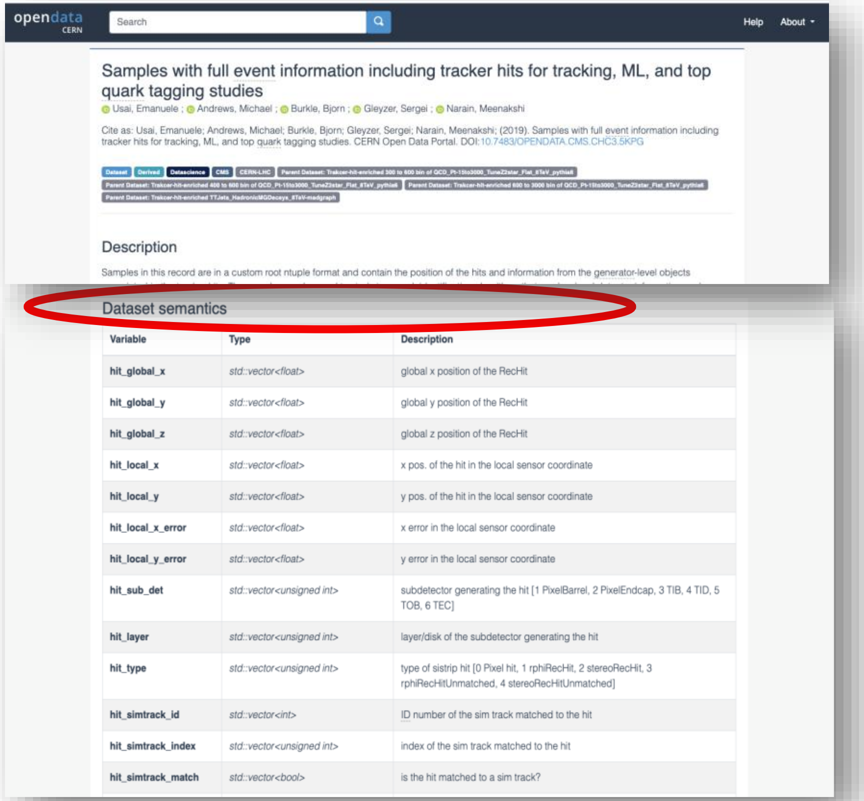
Data usually need to be integrated with other data, and with applications/workflows for analyis, storage and processing.

- Principles:
  - I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
  - I2: (Meta)data use vocabularies that follow FAIR principles
  - I3: (Meta)data include qualified references to other (meta)data

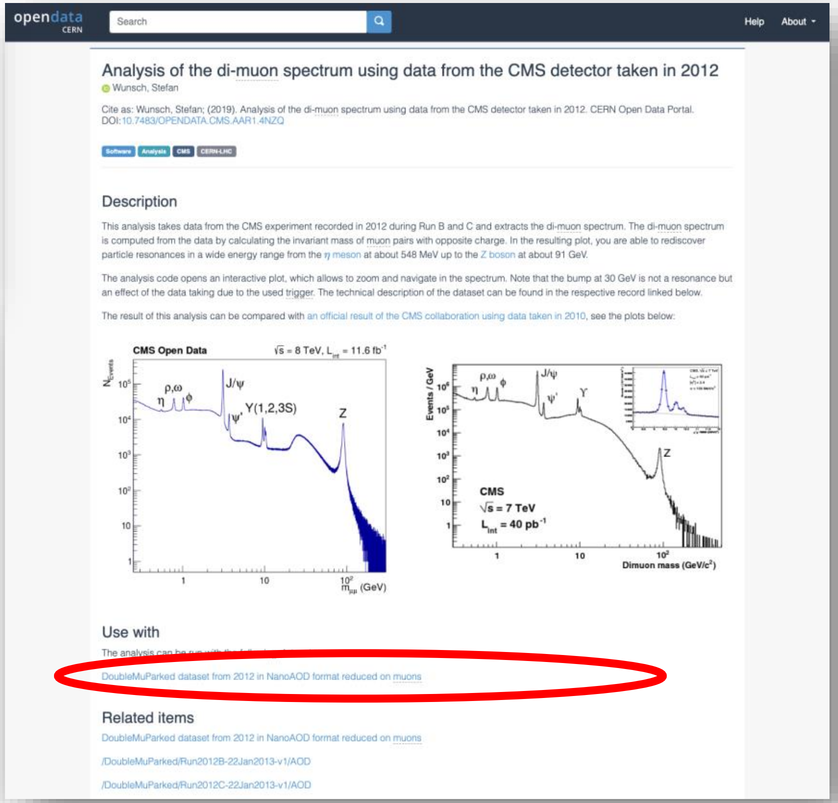
# COD: Interoperable




I1: Multiple data formats and common classifications




I2: Semantic descriptions



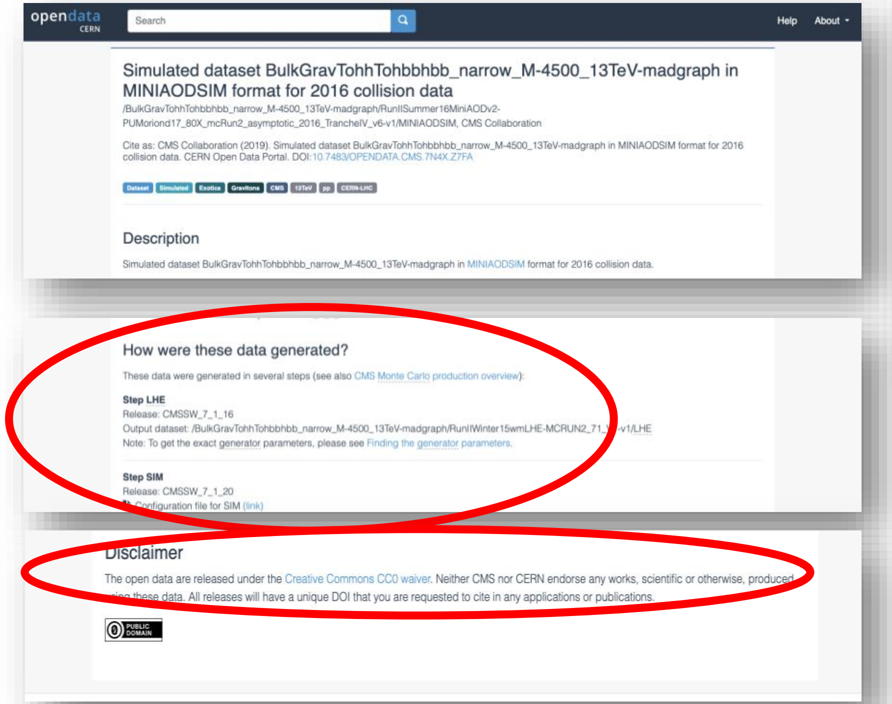

I3: Fully qualified references

# FAIR: Reusable

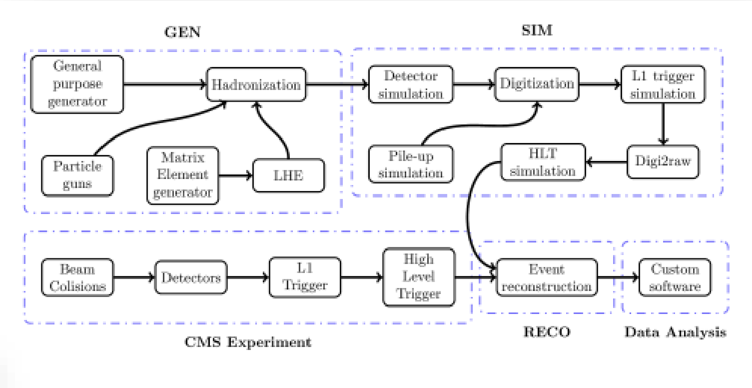
Metadata and data should be well-described so that they cn be replicated and/or combined in different settings

- Principles:
  - R1: (Meta)data are richly described with a plurality of accurate and relevant attributes
    - R1.1: (Meta)data are released with a clear and accessible data usage license
    - R1.2: (Meta)data are associated with detailed provenance
    - R1.3: (Meta)data meet domain-relevant community standards

# COD: Reusable

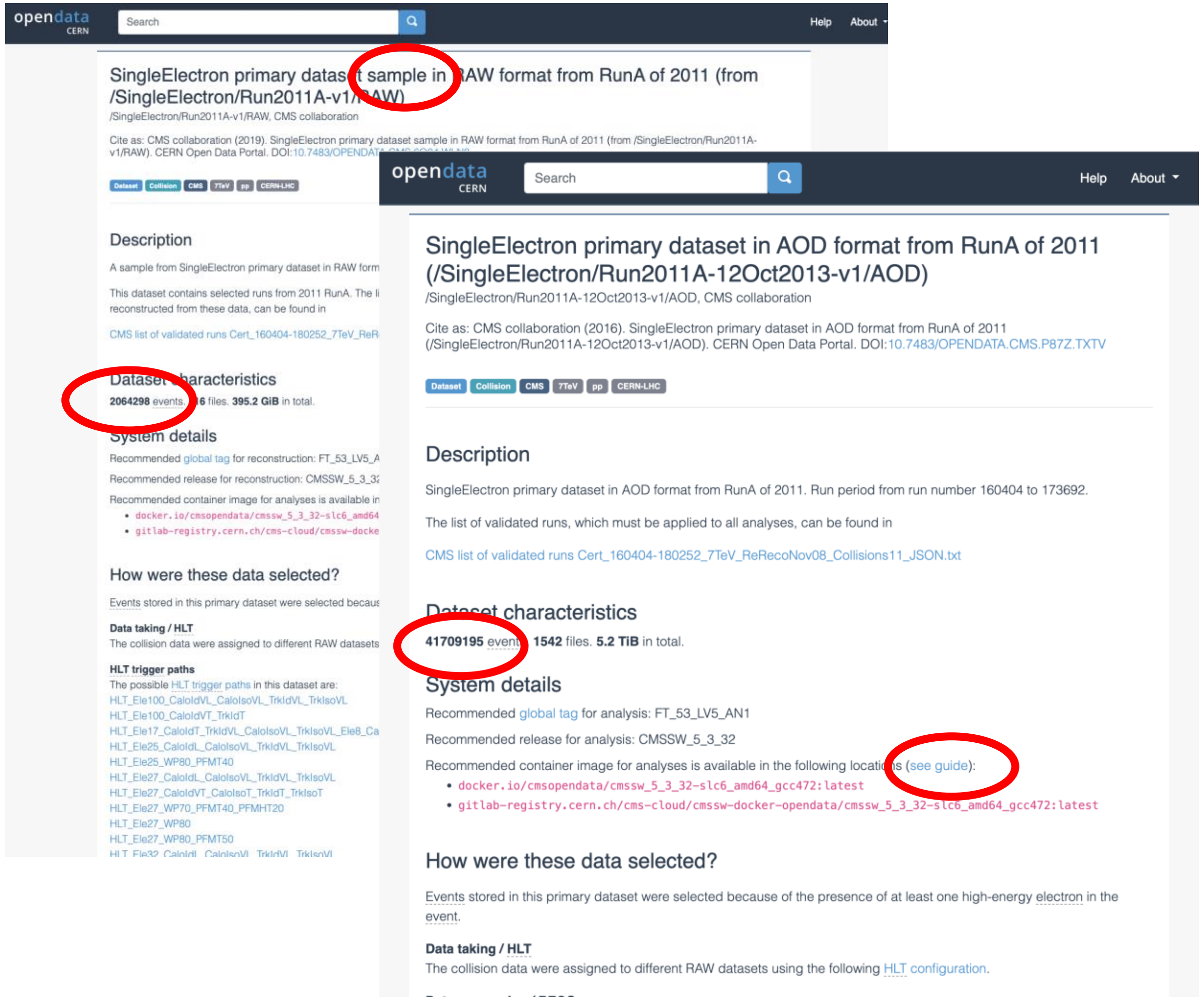


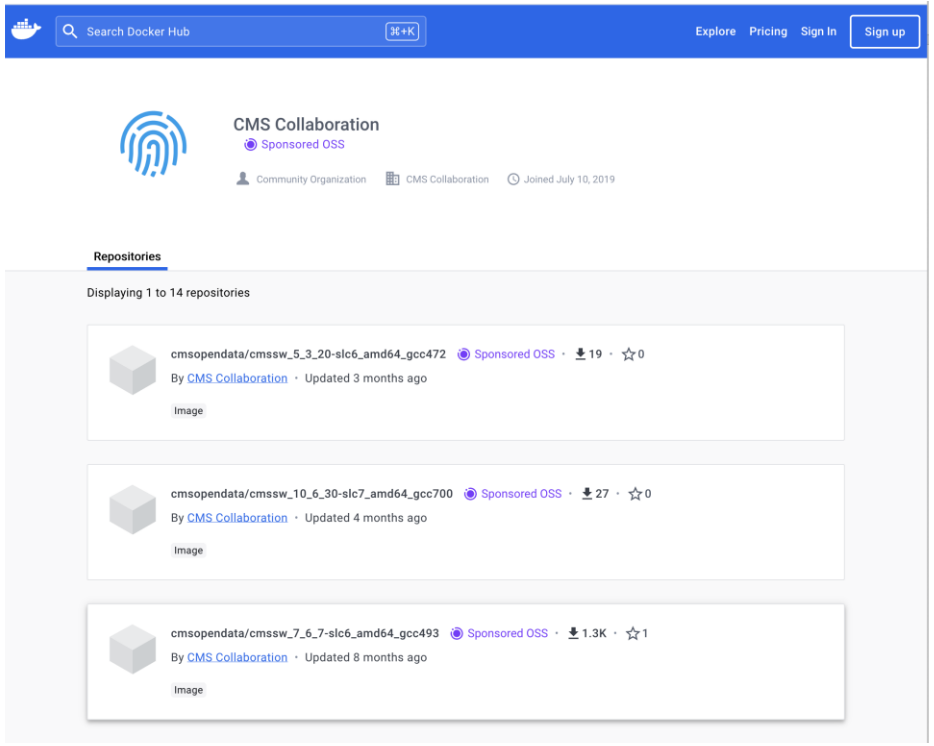

R1.1: Data usage license




R1.2: Provenance

# Reuse/reproduce: Can we reproduce analysis? sample?



# Reuse/reproduce: Can we reproduce analysis?

https://pubmed.ncbi.nlm.nih.gov/22675527/

> PLoS One. 2012;7(6):e38234. doi: 10.1371/journal.pone.0038234. Epub 2012 Jun 1.

## The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements

Ed H B M Gronenschild [1], Petra Habets, Heidi I L Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

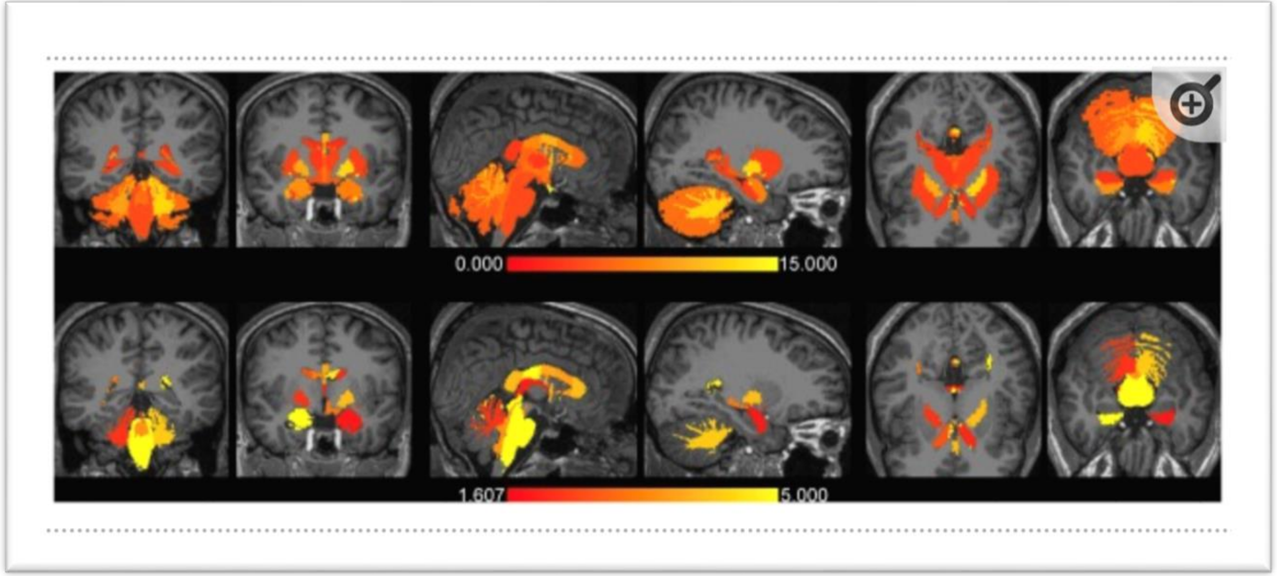
Affiliations + expand

PMID: 22675527 PMCID: PMC3365894 DOI: 10.1371/journal.pone.0038234

Free PMC article

### Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0




Software changes (Freesurfer 4.3.1, 4.5.0, 5.0.0): 8.8±6.6% (volume) and 2.8±1.3% (thickness)

Operating system changes (macOS 10.5, 10.6): about factor two smaller

# Four pillars of reusable computational research

I. **Input Data:**
   I. Input Files and parameters

II. **Analysis code:**
   I. User code
   II. Software frameworks

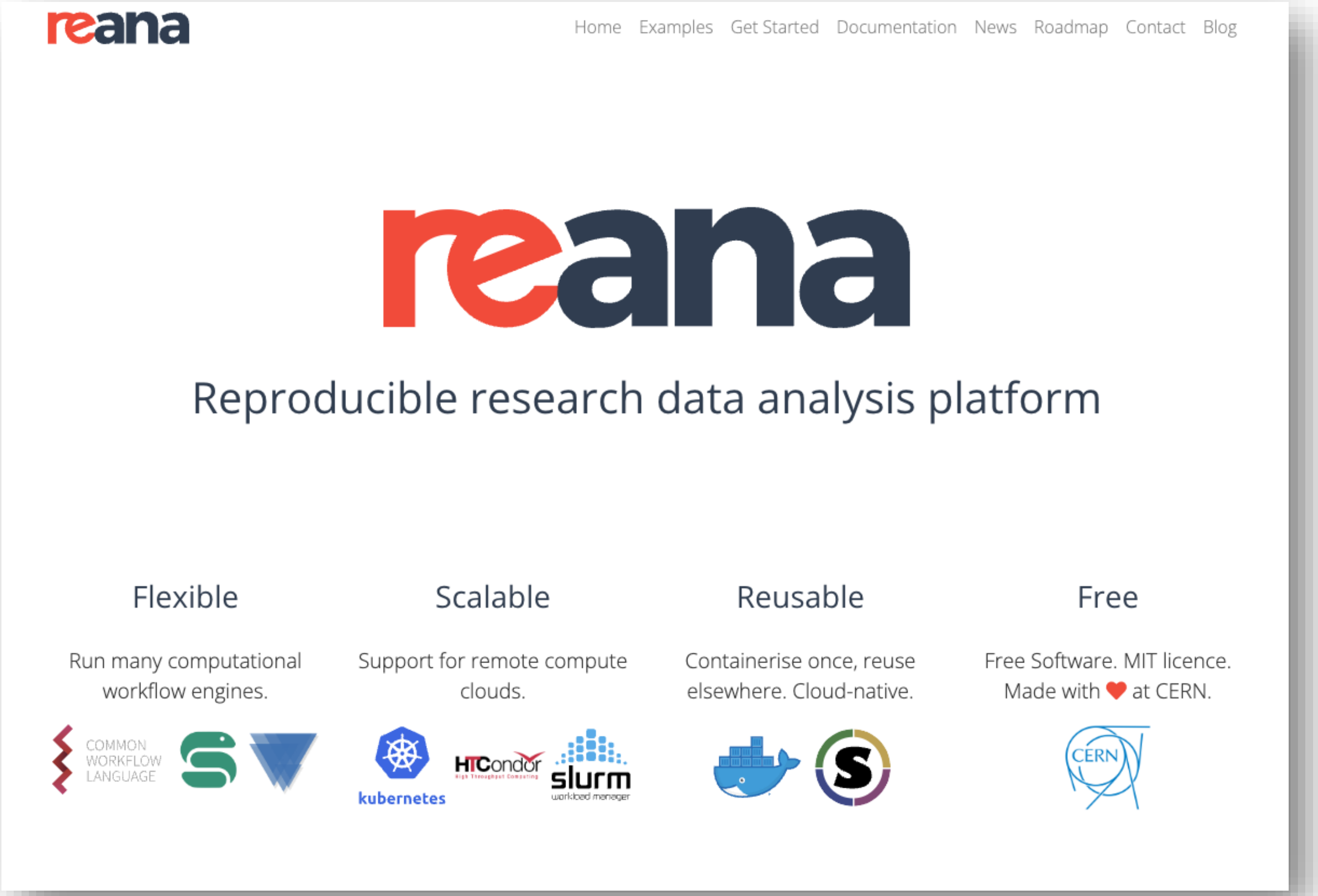
III. **Computing Environment:**
   I. Operating system
   II. Databases

IV. **Computational recipes:**
   I. Extra shell commands steps
   II. Notebooks and workflows

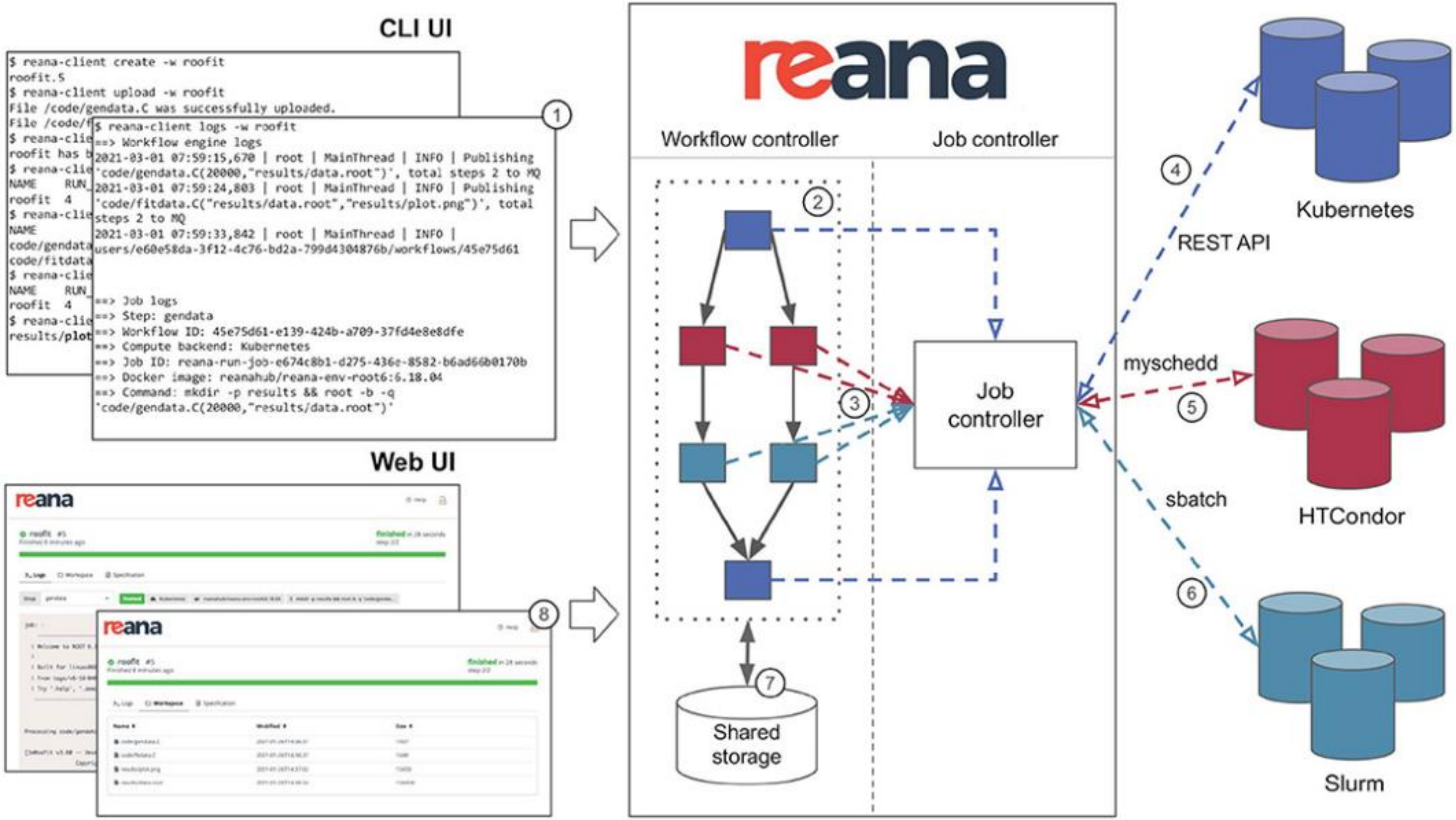
# REANA: Reusable analysis

http://reana.io

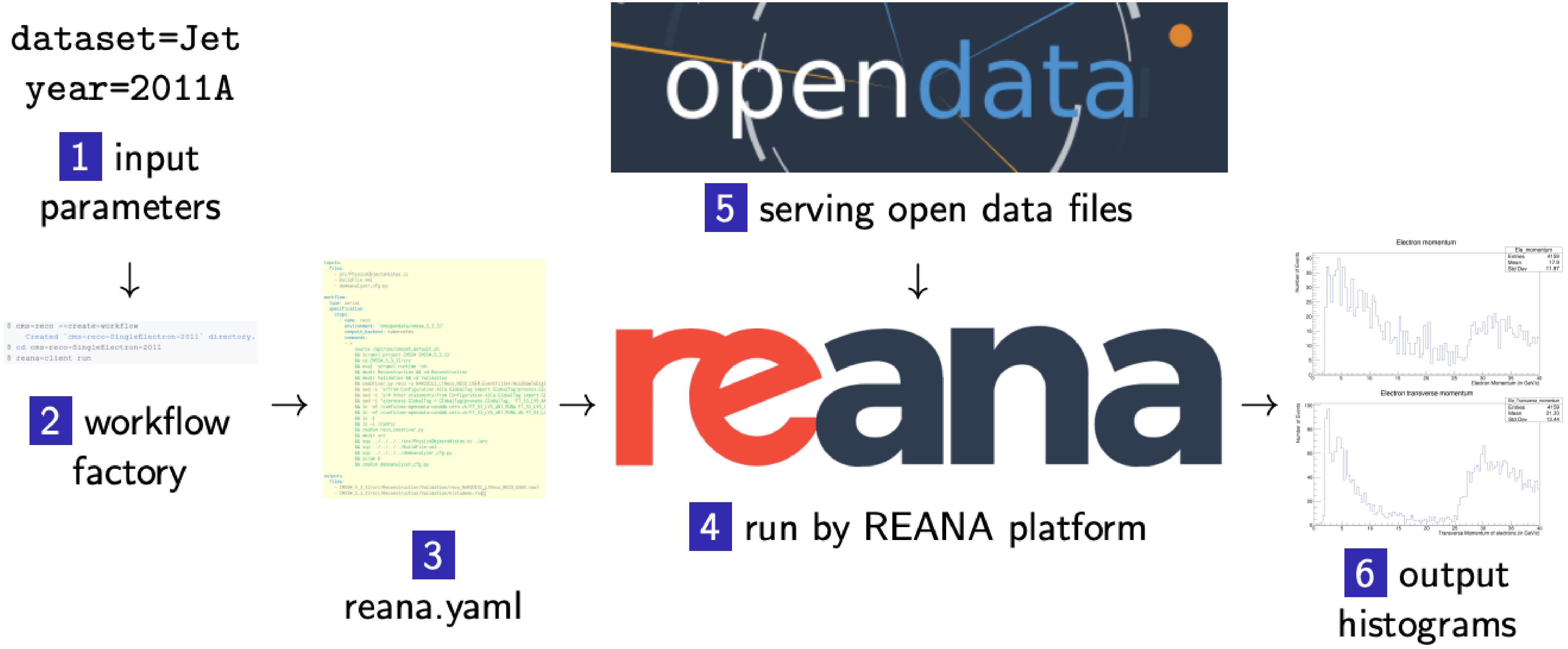
Deploy and run containerised workflows on compute clouds



# REANA in a nutshell



# Reprocessing CMS datasets on REANA

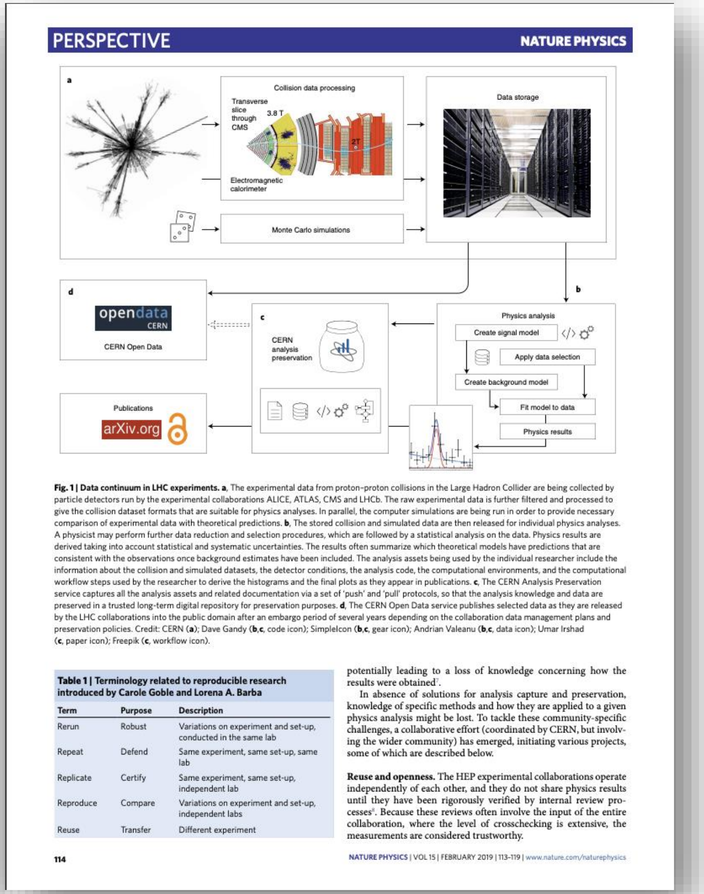



Parametrised workflow runnable on REANA reproducible analysis platform

# SUMMARY: Open is not enough

**F**indable
**A**ccessible
**I**nteroperable
**R**eusable

Data
+
Code
+
Environment
+
Workflow
=========
Reusability


https://www.nature.com/articles/s41567-018-0342-2.pdf home.cern