CERN Open Data Portal

Tibor Šimko

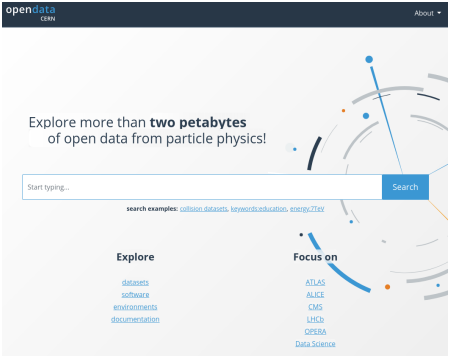
CERN

3rd DPHEP Collaboration Workshop, 21–23 June 2021

https://indico.cern.ch/event/1043155

CERN Open Data portal

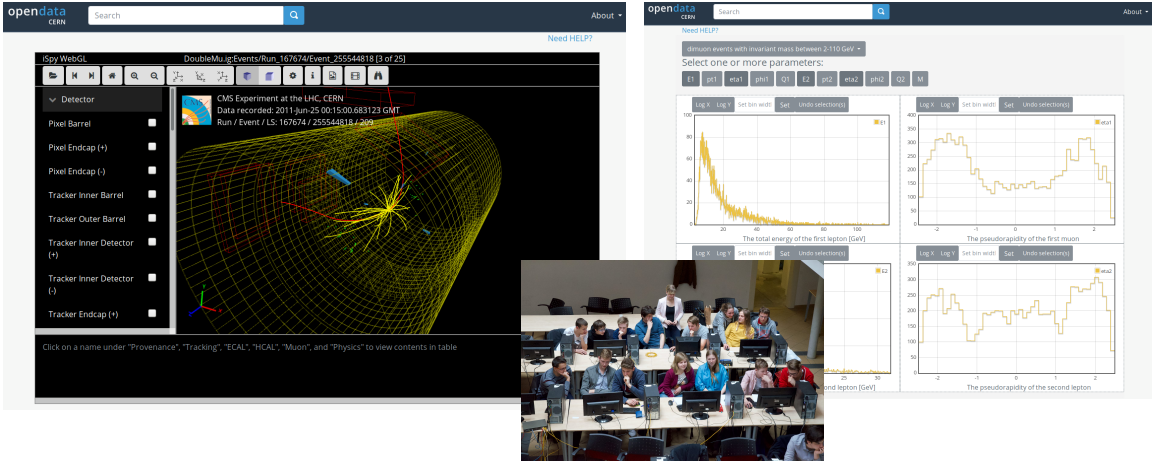
- launched in November 2014
- rich content
  - derived datasets for education
  - collision and simulated datasets for research
  - configuration files and documentation
  - virtual machines and container images
  - software tools and analysis examples
- total size in June 2021
  - over 7'600 bibliographic records
  - over 900'000 files
  - over 2.4 petabytes



https://opendata.cern.ch

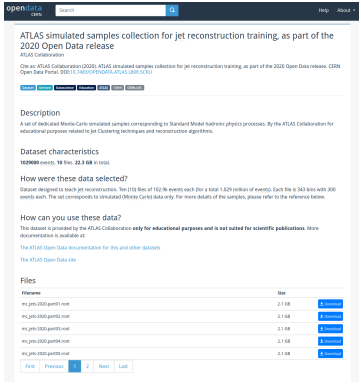
Developed by CERN in close collaboration with Experiments

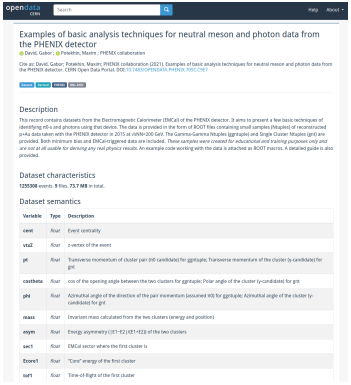
Education-oriented use cases



Interactive event display and histogramming for derived datasets

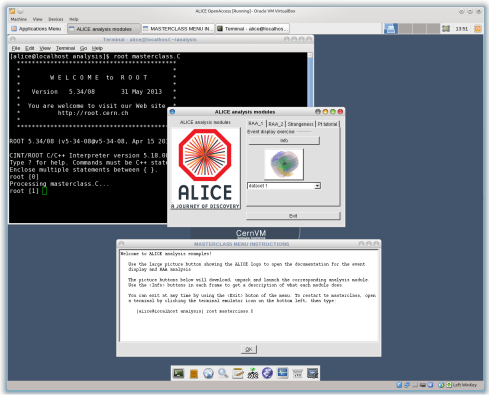
Latest education-level content news



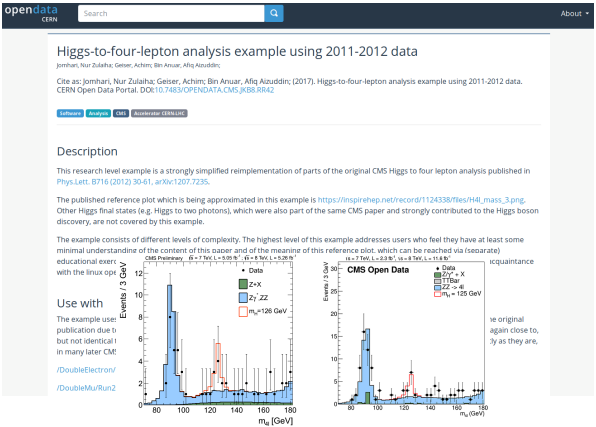


- **May 2021** ATLAS releases jet reconstruction training samples
- **June 2021** First open data release from the PHENIX experiment

Research-oriented use cases



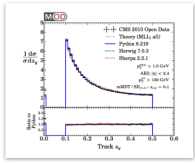
Run CernVM Virtual Machines

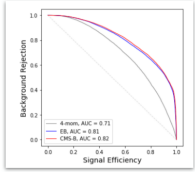


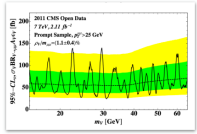
Run simplified research-level analysis examples

. . . enables independent theoretical research

Over forty papers citing CMS open data







arXiv:1704.05066 arXiv:1807.11916 arXiv:1902.04222

Searches, QCD jet studies, Machine Learning. . .

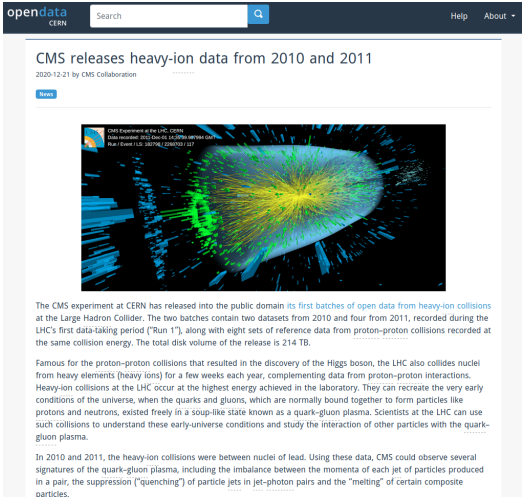
A measurement of the z_g distribution in pp collisions, using CMS open data, was recently reported [33, 34]. In PbPb collisions, this measurement reflects how the two color-charged par-

[33] A. Larkoski et al., "Exposing the QCD splitting function with CMS Open Data", *Phys. Rev. Lett.* **119** (2017) 132003, doi:10.1103/PhysRevLett.119.132003, arXiv:1704.05066.

[34] A. Tripathee et al., "Jet Substructure Studies with CMS Open Data", *Phys. Rev. D* **96** (2017) 074003, doi:10.1103/PhysRevD.96.074003, arXiv:1704.05842.

. . . that the CMS collaboration cites!

Latest research-level content news



- **December 2020** CMS released first 2010–11 heavy-ion data samples and reference proton-proton datasets (214 TB).

  `https://opendata.cern.ch/docs/cms-releases-heavy-ion-data`

- **October 2020** CMS run a CMS Open Data Workshop for Theorists at the LPC.

  `https://indico.cern.ch/event/882586/`

- **August 2020** CMS released fifth batch of new open data. All proton-proton collision data recorded in 2010–11 and half of 2012 are available throughout the portal.

  `https://opendata.cern.ch/docs/cms-completes-2010-2011-pp-data`

FAIR guiding principles for scientific data management

scientific **data**

Explore Content | Journal Information | Publish With Us

nature > scientific data > comment > article

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons

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

180k Accesses | 2306 Citations | 1797 Altmetric | Metrics

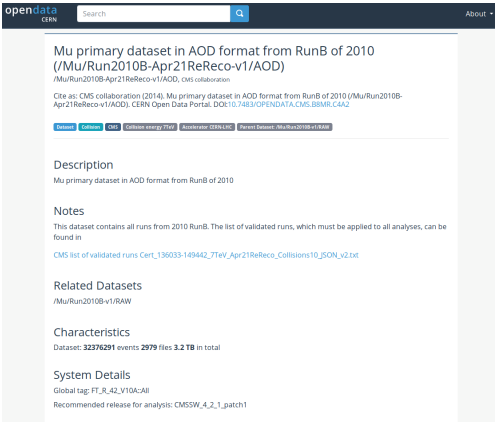
An Addendum to this article was published on 19 March 2019

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.

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

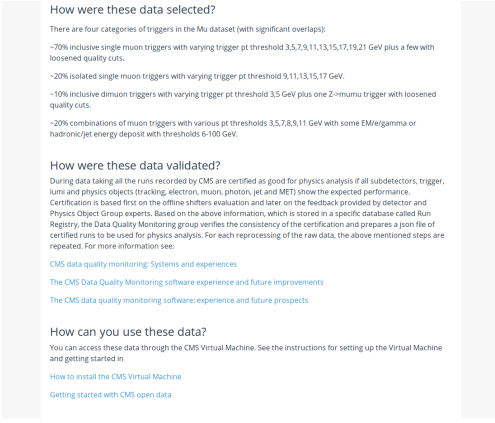
Findable: persistent identifiers & machine readability



`https://opendata.cern.ch/record/14` ≡ `http://doi.org/10.7483/OPENDATA.CMS.B8MR.C4A2`

Each dataset is identified by a “record ID” and optionally minted with a DOI

Findable: rich context description & machine readability



How were these data selected?

There are four categories of triggers in the Mu dataset (with significant overlaps):

~70% inclusive single muon triggers with varying trigger pt threshold 3,5,7,9,11,13,15,17,19,21 GeV plus a few with loosened quality cuts.

~20% isolated single muon triggers with varying trigger pt threshold 9,11,13,15,17 GeV.

~10% inclusive dimuon triggers with varying trigger pt threshold 3,5 GeV plus one Z->mumu trigger with loosened quality cuts.

~20% combinations of muon triggers with various pt thresholds 3,5,7,8,9,11 GeV with some EM/e/gamma or hadronic/jet energy deposit with thresholds 6-100 GeV.

How were these data validated?

During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and physics objects (tracking, electron, muon, photon, jet and MET) show the expected performance. Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:

CMS data quality monitoring: Systems and experiences

The CMS Data Quality Monitoring software experience and future improvements

The CMS data quality monitoring software: experience and future prospects

How can you use these data?

You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in

How to install the CMS Virtual Machine

Getting started with CMS open data

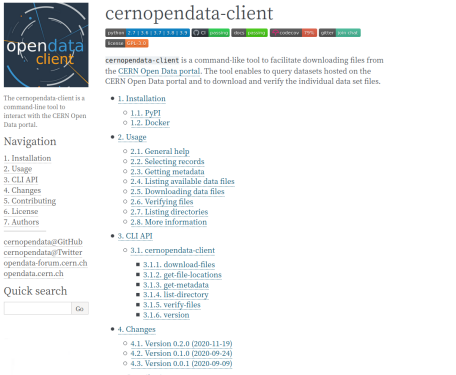

```
      ...
    }
  ],
  "run_period": [
    "Run2010B"
  ],
  "system_details": {
    "global_tag": "FT_R_42_V10A::All",
    "release": "CMSSW_4_2_8"
  },
  "title": "/Mu/Run2010B-Apr21ReReco-v1/AOD",
  "title_additional": "Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD)",
  "type": {
    "primary": "Dataset",
    "secondary": [
      "Collision"
    ]
  },
  "usage": {
    "description": "You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting
started in",
    "links": [
      {
        "description": "How to install the CMS Virtual Machine",
        "url": "/docs/cms-virtual-machine-2010"
      },
      {
        "description": "Getting started with CMS open data",
        "url": "/docs/cms-getting-started-2010"
      }
    ]
  },
  "validation": {
    "description": "During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and
physics objects (tracking, electron, muon, photon, jet and MET) show the expected performance. Certification is based first on the offline shifters
evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a
specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of
certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information
see:",
    "links": [
      {
        "description": "CMS data quality monitoring: Systems and experiences",
        "url": "http://iopscience.iop.org/1742-6596/219/7/072020/pdf/1742-6596_219_7_072020.pdf"
      },
      {
        "description": "The CMS Data Quality Monitoring software experience and future improvements",
        "url": "http://cds.cern.ch/record/1631039/files/CR2013_418.pdf"
      },
      {
        "description": "The CMS data quality monitoring software: experience and future prospects",
        "url": "http://iopscience.iop.org/1742-6596/513/3/032024/pdf/1742-6596_513_3_032024.pdf"
      }
    ]
  }
},
"updated": "2020-12-21T18:56:54.170861+00:00"
}
```


`https://opendata.cern.ch/record/14` $\equiv$ `http://doi.org/10.7483/OPENDATA.CMS.B8MR.C4A2`

Rich curated context information on data selection, validation, and use

Accessible: downloading content

Browsing and downloading files manually



- **September 2020** Automated client

Accessible: working with large datasets

A dataset of 9.6 TB size, 26'958 files

```
/tmp $ cernopendata-client download-files --help
Usage: cernopendata-client download-files [OPTIONS]

  Download data files belonging to a record.

  Select a CERN Open Data bibliographic record by a record ID, a DOI, or a
  title and download data files belonging to this record.

  Examples:
    $ cernopendata-client download-files --recid 5500
    $ cernopendata-client download-files --recid 5500 --filter-name BuildFile.xml
    $ cernopendata-client download-files --recid 5500 --filter-regexp py$
    $ cernopendata-client download-files --recid 5500 --filter-range 1-4
    $ cernopendata-client download-files --recid 5500 --filter-range 1-2,5-7
    $ cernopendata-client download-files --recid 5500 --filter-regexp py --filter-range 1-2
```

```
/tmp 1m 4s $ cernopendata-client get-file-locations --recid 8884 | head -3
http://opendata.cern.ch/eos/opendata/cms/MonteCarlo2012/Summer12_DR53X/QCD_Pt_
v1/00000/0052F9A9-5EB2-E311-81A9-0026189437ED.root
http://opendata.cern.ch/eos/opendata/cms/MonteCarlo2012/Summer12_DR53X/QCD_Pt_
v1/00000/00BF7DF8-69B2-E311-89E6-003048678E8A.root
http://opendata.cern.ch/eos/opendata/cms/MonteCarlo2012/Summer12_DR53X/QCD_Pt_
v1/00000/020194EC-CAB2-E311-8B3A-003048678B94.root
```

Downloading file locations

```
/tmp $ cernopendata-client download-files --recid 8884 --filter-range 1
==> Downloading file 1 of 1
  -> File: ./8884/0052F9A9-5EB2-E311-81A9-0026189437ED.root
  -> Progress: 38869/150643 KiB (25%)
```

Downloading and verifying files

Interoperable: data semantics

Dataset characteristics

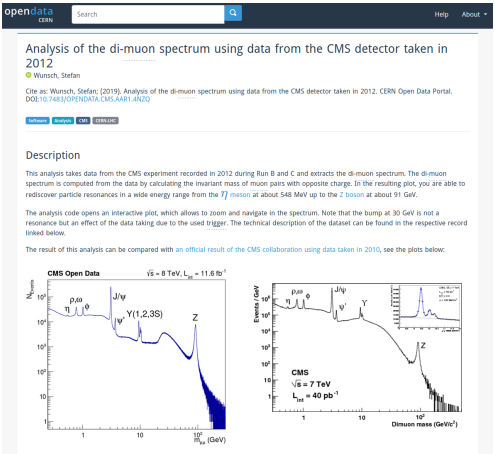
9430509 events. **14825** files. **12.8 TB** in total.

Dataset semantics

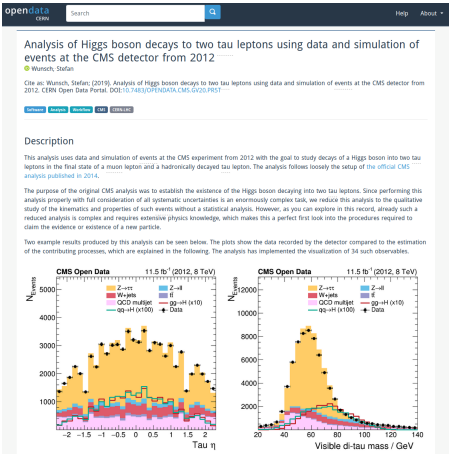
Variable	Type	Description
hit_global_x	*std::vector<float>*	global x position of the RecHit
hit_global_y	*std::vector<float>*	global y position of the RecHit
hit_global_z	*std::vector<float>*	global z position of the RecHit
hit_local_x	*std::vector<float>*	x pos. of the hit in the local sensor coordinate
hit_local_y	*std::vector<float>*	y pos. of the hit in the local sensor coordinate
hit_local_x_error	*std::vector<float>*	x error in the local sensor coordinate
hit_local_y_error	*std::vector<float>*	y error in the local sensor coordinate
hit_sub_det	*std::vector<unsigned int>*	subdetector generating the hit [1 PixelBarrel, 2 PixelEndcap, 3 TIB, 4 TID, 5 TOB, 6 TEC]
hit_layer	*std::vector<unsigned int>*	layer/disk of the subdetector generating the hit
hit_type	*std::vector<unsigned int>*	type of sistrip hit [0 Pixel hit, 1 rphiRecHit, 2 stereoRecHit, 3 rphiRecHitUnmatched, 4 stereoRecHitUnmatched]
hit_simtrack_id	*std::vector<int>*	ID number of the sim track matched to the hit
hit_simtrack_index	*std::vector<unsigned int>*	index of the sim track matched to the hit
hit_simtrack_match	*std::vector<bool>*	is the hit matched to a sim track?
hit_genparticle_id	*std::vector<unsigned int>*	index of the gen particle matched to the hit
hit_pdgid	*std::vector<int>*	PDG ID of the gen particle matched to the hit
hit_recotrack_id	*std::vector<unsigned int>*	index of the reco track matched to the hit
hit_recotrack_match	*std::vector<bool>*	is the hit matched to a reco track?
hit_genparticle_match	*std::vector<bool>*	is the hit matched to a gen particle?
hit_genjet_id	*std::vector<unsigned int>*	index of the gen jet matched to the hit

Dataset variables coming with detailed semantics description

Interoperable: accompanying examples



Dimuon spectrum analysis example



$H \to \tau\tau$ analysis example

Reusable: how were the data generated?

Simulated dataset BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph in MINIAODSIM format for 2016 collision data

/BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM, CMS Collaboration

Cite as: CMS Collaboration (2019). Simulated dataset BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph in MINIAODSIM format for 2016 collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.7N4X.Z7FA

Dataset | Simulated | Exotica | Gravitons | CMS | 13TeV | CERN-LHC

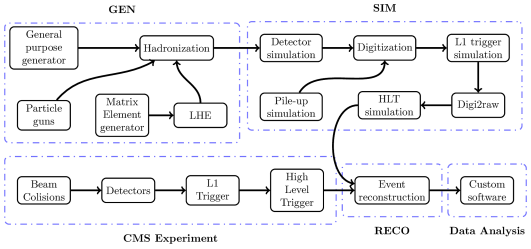
How were these data generated?

These data were generated in several steps (see also CMS Monte Carlo production overview):

Step LHE
Release: CMSSW_7_1_16
Output dataset: /BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIIWinter15wmLHE-MCRUN2_71_V1-v1/LHE
Note: To get the exact generator parameters, please see Finding the generator parameters.

Step SIM
Release: CMSSW_7_1_20
Configuration file for SIM (link)
Output dataset: /BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer15GS-MCRUN2_71_V1-v1/GEN-SIM

Step HLT RECO
Release: CMSSW_8_0_21



- full capture of data generation steps
- full capture of compute environments
- full capture of configuration files
- full capture of production scripts

Data records come with full provenance information

Reusable: can we reprocess published data samples?

SingleElectron primary dataset sample in RAW format from RunA of 2011 (from /SingleElectron/Run2011A-v1/RAW)

/SingleElectron/Run2011A-v1/RAW, CMS collaboration

Cite as: CMS collaboration (2019). SingleElectron primary dataset sample in RAW format from RunA of 2011 (from /SingleElectron/Run2011A-v1/RAW). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.6O84.WLN8

Dataset | Collision | CMS | 7TeV | CERN-LHC

Description

A sample from SingleElectron primary dataset in RAW format from RunA of 2011. Run range [161224,163286].

This dataset contains selected runs from 2011 RunA. The list of validated lumi sections, which must be applied to all analyses on events reconstructed from these data, can be found in

CMS list of validated runs Cert_160404-180252_7TeV_ReRecoNov08_Collisions11_JSON.txt

Dataset characteristics

2064298 events. **116** files. **424.3 GB** in total.

How can you use these data?

These data are in RAW format and not directly usable in analysis. The reconstructed data reprocessed from these RAW data are included in the data of this record. The reconstruction step can be repeated with the configuration file below and the resulting AOD has been confirmed to be identical with the original one with comparison code available in Validation code to plot basic physics objects from AOD

RAW

SingleElectron primary dataset in AOD format from RunA of 2011 (/SingleElectron/Run2011A-12Oct2013-v1/AOD)

/SingleElectron/Run2011A-12Oct2013-v1/AOD, CMS collaboration

Cite as: CMS collaboration (2016). SingleElectron primary dataset in AOD format from RunA of 2011 (/SingleElectron/Run2011A-12Oct2013-v1/AOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.P87Z.TXTV

Dataset | Collision | CMS | 7TeV | CERN-LHC

Description

SingleElectron primary dataset in AOD format from RunA of 2011. Run period from run number 160404 to 173692.

This dataset contains all runs from 2011 RunA. The list of validated runs, which must be applied to all analyses, can be found in

CMS list of validated runs Cert_160404-180252_7TeV_ReRecoNov08_Collisions11_JSON.txt

Dataset characteristics

41709195 events. **1542** files. **5.8 TB** in total.

How were these data selected?

Events stored in this primary dataset were selected because of the presence of at least one high-energy electron 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: CMSSW_5_3_12_patch1
Global tag: FT_R_53_LV5::All
Configuration file for RECO step reco_2011A_SingleElectron

AOD

Reprocessing workflow: computational recipe

3. Workflow

The workflow can be logically divided into several parts:

0. ***Upload all files.***
   Some files cannot be generated at run time and need to be uploaded.

```
inputs:
  files:
    - src/PhysicsObjectsHistos.cc
    - BuildFile.xml
    - demoanalyzer_cfg.py
```

1. ***Fix the CMS SW environment variables manually.***
   First, we have to set up the environment variables accordingly for the CMS SW. Although this is done in the docker image, reana overrides them and they need to be reset. This is done by invoking the cms entrypoint.sh script commands.

   See also this issue.

```
$ source /opt/cms/cmsset_default.sh
$ scramv1 project CMSSW CMSSW_5_3_32
$ cd CMSSW_5_3_32/src
$ eval `scramv1 runtime -sh`
```

2. ***Create the specific CMS path.***
   CMS specific data analysis framework requires two directory levels. See also this issue.

```
$ mkdir Reconstruction && cd Reconstruction
$ mkdir Validation && cd Validation
```

3. ***Create the reconstruction file.***
   See also this repo.

```
$ cmsDriver.py reco -s RAW2DIGI,L1Reco,RECO,USER:EventFilter/HcalRawToDigi/hcallaserhbhehffilter2012_cf
```

4. ***Adjust the reconstruction file to the specific data file.***
   Although generated using parameters, the reconstruction file still requires changes.

```
$ sed -i 's/from Configuration.AlCa.GlobalTag import GlobalTag/process.GlobalTag.connect = cms.string("
$ sed -i 's/# Other statements/from Configuration.AlCa.GlobalTag import GlobalTag/g' reco_cmsdriver.py
$ sed -i "s/process.GlobalTag = GlobalTag(process.GlobalTag, 'FT_53_LV5_AN1::All', '')/process.GlobalTa
```

5. ***Link the CVMFS files.***
   The ls -l commands are explicitly needed to make sure that the cms-opendata-conddb.cern.ch directory has actually expanded in the image, according to this guide. See also this issue.

```
$ ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA FT_53_LV5_AN1
$ ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA.db FT_53_LV5_AN1_RUNA.db
$ ls -l
$ ls -l /cvmfs/
```

6. ***Run the reconstruction.***
   At this point all environment variables and files should be proper.

```
$ cmsRun reco_cmsdriver.py
```

7. ***Adjust project structure for validation***
   Copy the required files for the next steps.

```
$ mkdir src
$ scp ../../../../src/PhysicsObjectsHistos.cc ./src
$ scp ../../../../BuildFile.xml .
$ scp ../../../../demoanalyzer_cfg.py .
```

8. ***Run CMS scram command to fix libraries.***
   Most importantly, the *BuildFile.xml* has to be inside the directory where the *scram* command is executed.

```
$ scram b
```

9. ***Run the validation file.***
   See also this repo

```
$ cmsRun demoanalyzer_cfg.py
```

Workflow steps to run CMS reconstruction in CMSSW environment

Reprocessing CMS datasets on REANA

```
dataset=Jet
year=2011A
```

1 input parameters

↓

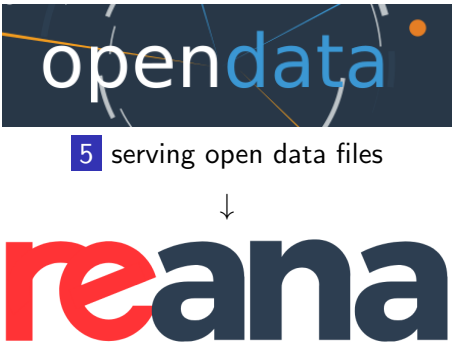
```
$ cms-reco --create-workflow
    Created 'cms-reco-SingleElectron-2011' directory.
$ cd cms-reco-SingleElectron-2011
$ reana-client run
```

2 workflow factory

→

3 reana.yaml

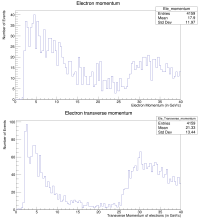
→



5 serving open data files

4 run by REANA platform

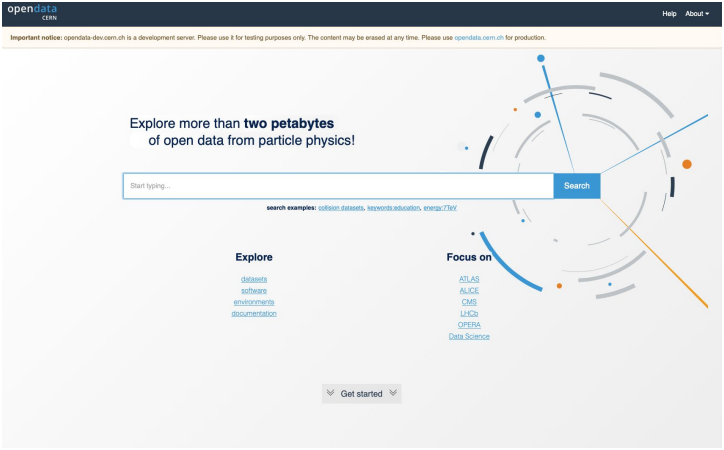
→



6 output histograms

Parametrised workflow runnable on REANA reproducible analysis platform

Latest technology news



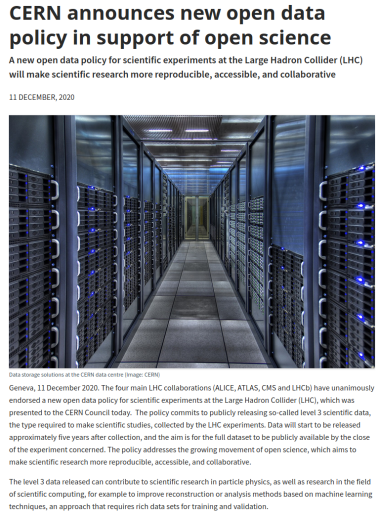
- **March 2021**: Major upgrade of the underlying digital repository software framework (Invenio 3.4, Elasticsearch 7, Angular → React, Bootstrap → SemanticUI, . . .)

Future plans

- **December 2020**: A common statement on the open data policy by CERN management and ATLAS, ALICE, CMS, LHCb and TOTEM experiments.

  `https://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments`

- Prepare for forthcoming increase in open data publishing.
- Introduce flexible hot/cold disk/tape storage solution. Part of dataset files on disk, part on tapes.
- Simplify ingestion and exposure of experiment datasets (Rucio, Dirac).
- Further automatise provenance testing and usage examples.



CERN announces new open data policy in support of open science

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

11 DECEMBER, 2020

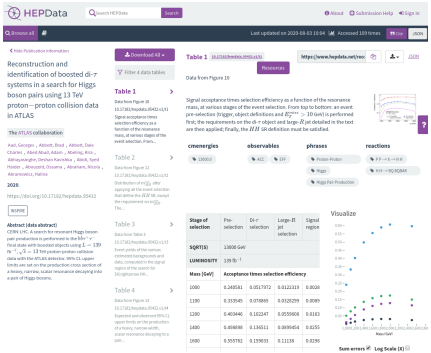
Data storage solutions at the CERN data centre (Image: CERN)

Geneva, 11 December 2020. The four main LHC collaborations (ALICE, ATLAS, CMS and LHCb) have unanimously endorsed a new open data policy for scientific experiments at the Large Hadron Collider (LHC), which was presented to the CERN Council today. The policy commits to publicly releasing so-called level 3 scientific data, the type required to make scientific studies, collected by the LHC experiments. Data will start to be released approximately five years after collection, and the aim is for the full dataset to be publicly available by the close of the experiment concerned. The policy addresses the growing movement of open science, which aims to make scientific research more reproducible, accessible, and collaborative.

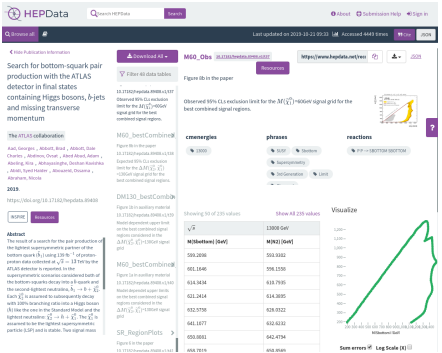
The level 3 data released can contribute to scientific research in particle physics, as well as research in the field of scientific computing, for example to improve reconstruction or analysis methods based on machine learning techniques, an approach that requires rich data sets for training and validation.

An ecosystem of sister data repositories

HEPData widens publication-level data scope



ATLAS collaboration `https://www.hepdata.net/record/ins1809175`

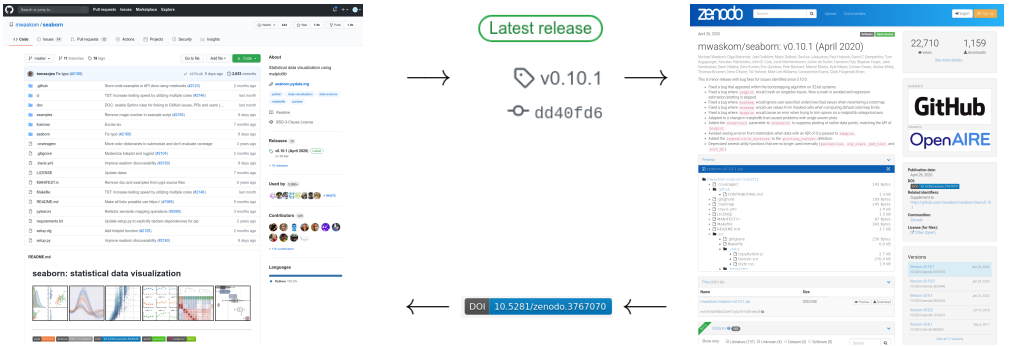


ATLAS collaboration `https://www.hepdata.net/record/ins1748602`

HEPData provides interactive interface to explore and download publication-level data behind plots and tables

. . . now starting to handle more data types: likelihoods!

Zenodo preserves research software

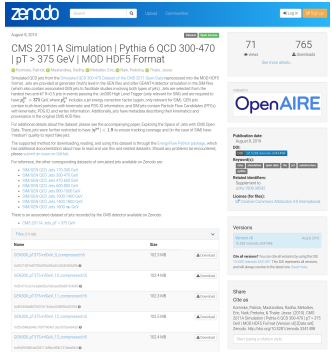


https://guides.github.com/activities/citable-code

GitHub ↔ Zenodo bridge to automatically preserve releases

Zenodo hosts Machine Learning community data

Di Florio *et al* `http://opendata.cern.ch/record/12320`

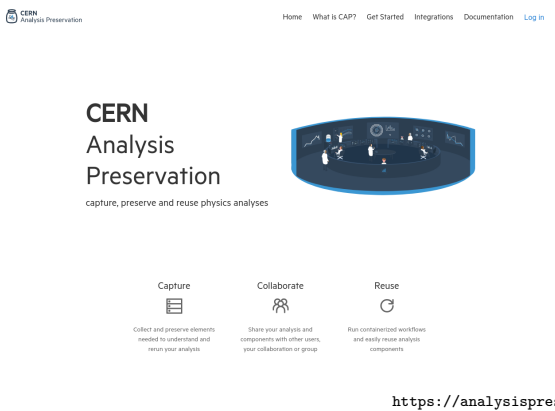


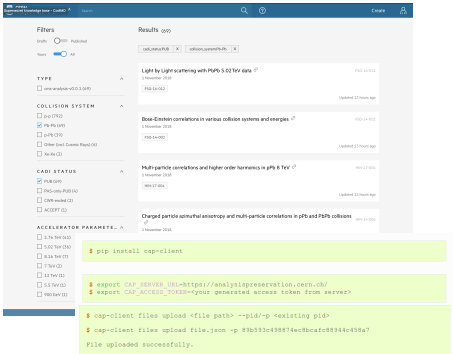
Komiske *et al* `https://zenodo.org/record/3341498`

Machine-learning related datasets on the CERN Open Data portal

Machine-learning related datasets on Zenodo

Preserving “restricted” analysis knowledge

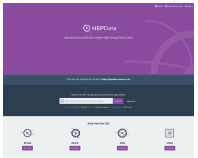




`https://analysispreservation.cern.ch`

CERN Analysis Preservation framework for collaboration-restricted data.
Following the same FAIR principles (FAIR $\neq$ open!)

A family of digital repositories in movement



publication-level data

experimental collaborations

invidivual researchers

event-level data

INVENIO