Open data provenance and reproducibility: a case study from publishing CMS open data

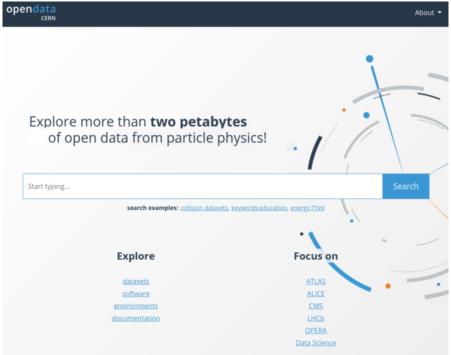
Tibor Šimko[1] Heitor Pascoal de Bittencourt[2] Edgar Carrera[2] Clemens Lange[2]
Kati Lassila-Perini[2] Lara Lloret[2] Tom McCauley[2] Jan Okraska[1]
Daniel Prelipcean[1] Mantas Savaniakas[2]
on behalf of the CERN Open Data team and the CMS Collaboration

[1] CERN Open Data team [2] CMS Collaboration

24th International Conference on Computing in High Energy and Nuclear Physics (CHEP)
Adelaide, Australia, 4–8 November 2019

CERN Open Data

- launched in November 2014
- rich content
  - collision and simulated datasets for research
  - derived datasets for education
  - configuration files and documentation
  - virtual machines and container images
  - software tools and analysis examples
- total size in November 2019
  - over 7'000 bibliographic records
  - over 800'000 files
  - over 2 petabytes



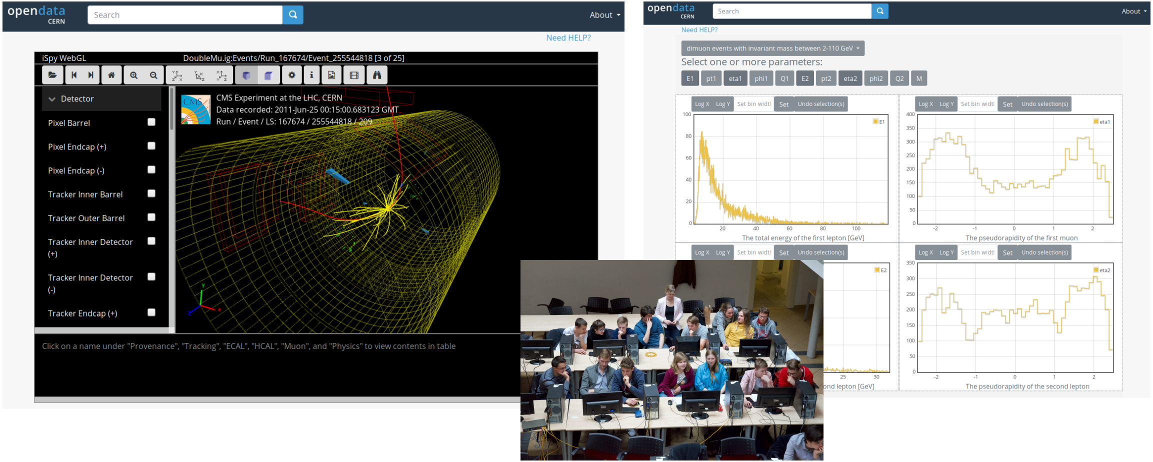
Explore more than **two petabytes** of open data from particle physics!

search examples: collision datasets, keywords:education, energy:7TeV

`http://opendata.cern.ch`

Developed by CERN-IT in close collaboration with Experiments

Education-oriented use cases

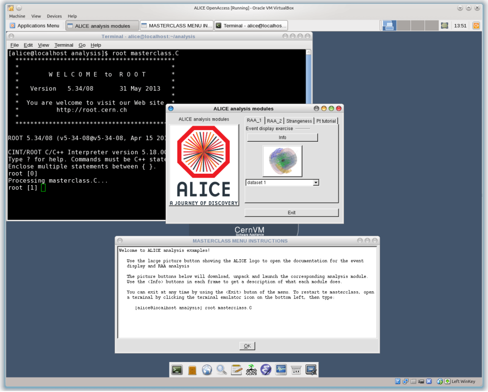


opendata CERN
Search
About
Need HELP?
iSpy WebGL
DoubleMu.ig:Events/Run_167674/Event_255544818 [3 of 25]
Detector
Pixel Barrel
Pixel Endcap (+)
Pixel Endcap (-)
Tracker Inner Barrel
Tracker Outer Barrel
Tracker Inner Detector (+)
Tracker Inner Detector (-)
Tracker Endcap (+)
CMS Experiment at the LHC, CERN
Data recorded: 2011-Jun-25 00:15:00.683123 GMT
Run / Event / LS: 167674 / 255544818 / 209
opendata CERN
Search
About
Need HELP?
dimuon events with invariant mass between 2-110 GeV
Select one or more parameters:
E1
pt1
eta1
phi1
Q1
E2
pt2
eta2
phi2
Q2
M
Log X
Log Y
Set bin widt
Set
Undo selection(s)
E1
The total energy of the first lepton [GeV]
eta1
The pseudorapidity of the first muon
E2
ond lepton [GeV]
eta2
The pseudorapidity of the second lepton

Click on a name under "Provenance", "Tracking", "ECAL", "HCAL", "Muon", and "Physics" to view contents in table

Interactive event display and histogramming for derived datasets

Research-oriented use cases

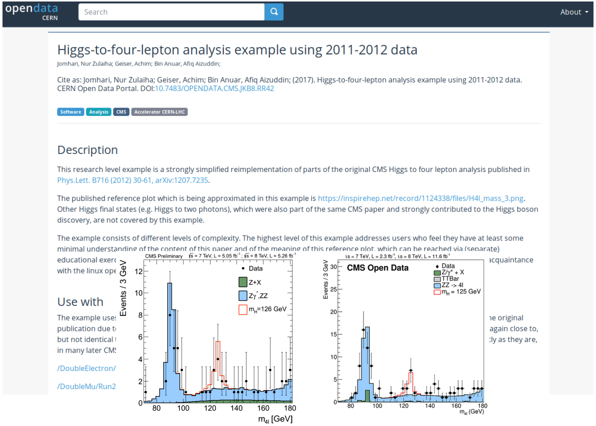


ALICE OpenAccess [Running] - Oracle VM VirtualBox
Machine View Devices Help
Applications Menu
ALICE analysis modules
MASTERCLASS MENU IN...
Terminal - alice@localhos...
13:51
Terminal - alice@localhost:~/analysis
File Edit View Terminal Go Help
ALICE analysis modules
RAA_1 RAA_2 Strangeness Pt tutorial
Event display exercise
Info
dataset 1
Exit
ALICE
A JOURNEY OF DISCOVERY
CernVM
Software Appliance
MASTERCLASS MENU INSTRUCTIONS
OK
Left WinKey

```
[alice@localhost analysis]$ root masterclass.C
  *******************************************
  *                                         *
  *        W E L C O M E  to  R O O T       *
  *                                         *
  *   Version   5.34/08       31 May 2013   *
  *                                         *
  *  You are welcome to visit our Web site  *
  *          http://root.cern.ch
  *
  *******************************************

ROOT 5.34/08 (v5-34-08@v5-34-08, Apr 15 201

CINT/ROOT C/C++ Interpreter version 5.18.00
Type ? for help. Commands must be C++ state
Enclose multiple statements between { }.
root [0]
Processing masterclass.C...
root [1]
```

```
Welcome to ALICE analysis examples!

   Use the large picture button showing the ALICE logo to open the documentation for the event
   display and RAA analysis

   The picture buttons below will download, unpack and launch the corresponding analysis module.
   Use the <Info> buttons in each frame to get a description of what each module does.

   You can exit at any time by using the <Exit> buton of the menu. To restart te masterclass, open
   a terminal by clicking the terminal emulator icon on the botton left, then type:

      [alice@localhost analysis] root masterclass.C
```

Run CernVM Virtual Machines



Higgs-to-four-lepton analysis example using 2011-2012 data

Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin;

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JKB8.RR42

Description

This research level example is a strongly simplified reimplementation of parts of the original CMS Higgs to four lepton analysis published in Phys.Lett. B716 (2012) 30-61, arXiv:1207.7235.

The published reference plot which is being approximated in this example is https://inspirehep.net/record/1124338/files/H4l_mass_3.png. Other Higgs final states (e.g. Higgs to two photons), which were also part of the same CMS paper and strongly contributed to the Higgs boson discovery, are not covered by this example.

The example consists of different levels of complexity. The highest level of this example addresses users who feel they have at least some minimal understanding of the content of this paper and of the meaning of this reference plot, which can be reached via (separate) educational exer acquaintance with the linux ope

Use with

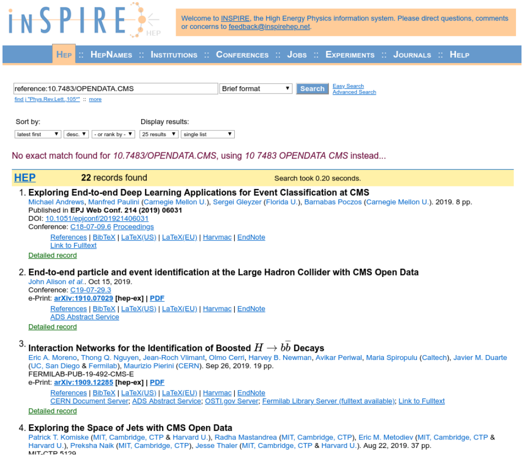
The example use he original publication due t again close to, but not identical tly as they are, in many later CM

/DoubleElectron/

/DoubleMu/Run2

Run realistic physics analysis examples

Enables independent theoretical research



Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

reference:10.7483/OPENDATA.CMS | Brief format | Search | Easy Search | Advanced Search

find j "Phys.Rev.Lett.,105*" :: more

Sort by: latest first | desc. | - or rank by -

Display results: 25 results | single list

No exact match found for *10.7483/OPENDATA.CMS*, using *10 7483 OPENDATA CMS* instead...

HEP 22 records found Search took 0.20 seconds.

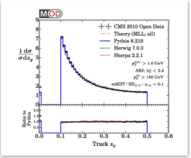
1. **Exploring End-to-end Deep Learning Applications for Event Classification at CMS**
Michael Andrews, Manfred Paulini (Carnegie Mellon U.), Sergei Gleyzer (Florida U.), Barnabas Poczos (Carnegie Mellon U.). 2019. 8 pp.
Published in **EPJ Web Conf. 214 (2019) 06031**
DOI: 10.1051/epjconf/201921406031
Conference: C18-07-09.6 Proceedings
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
Link to Fulltext
Detailed record

2. **End-to-end particle and event identification at the Large Hadron Collider with CMS Open Data**
John Alison *et al.*. Oct 15, 2019.
Conference: C19-07-29.3
e-Print: **arXiv:1910.07029 [hep-ex]** | **PDF**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
ADS Abstract Service
Detailed record

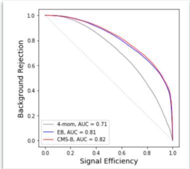
3. **Interaction Networks for the Identification of Boosted $H \to b\bar{b}$ Decays**
Eric A. Moreno, Thong Q. Nguyen, Jean-Roch Vlimant, Olmo Cerri, Harvey B. Newman, Avikar Periwal, Maria Spiropulu (Caltech), Javier M. Duarte (UC, San Diego & Fermilab), Maurizio Pierini (CERN). Sep 26, 2019. 19 pp.
FERMILAB-PUB-19-492-CMS-E
e-Print: **arXiv:1909.12285 [hep-ex]** | **PDF**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
CERN Document Server; ADS Abstract Service; OSTI.gov Server; Fermilab Library Server (fulltext available); Link to Fulltext
Detailed record

4. **Exploring the Space of Jets with CMS Open Data**
Patrick T. Komiske (MIT, Cambridge, CTP & Harvard U.), Radha Mastandrea (MIT, Cambridge, CTP), Eric M. Metodiev (MIT, Cambridge, CTP & Harvard U.), Preksha Naik (MIT, Cambridge, CTP), Jesse Thaler (MIT, Cambridge, CTP & Harvard U.). Aug 22, 2019. 37 pp.
MIT-CTP 5129

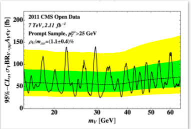
Over twenty papers citing CMS open data



arXiv:1704.05066

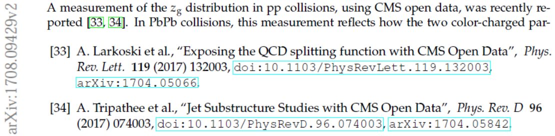


arXiv:1807.11916



arXiv:1902.04222

Searches, QCD jet studies, Machine Learning. . .



arXiv:1708.09429v2

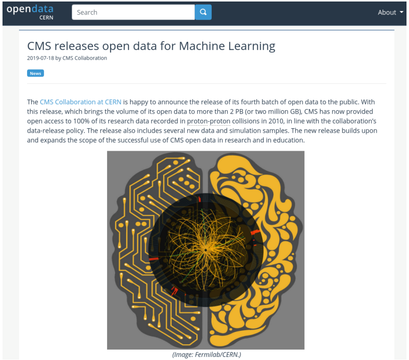
A measurement of the z_g distribution in pp collisions, using CMS open data, was recently reported [33, 34]. In PbPb collisions, this measurement reflects how the two color-charged par-

[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 start to cite!

New CMS open data release



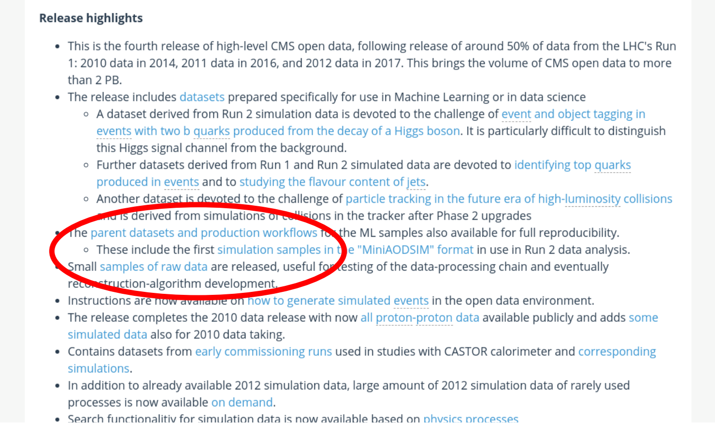
CMS releases open data for Machine Learning

2019-07-18 by CMS Collaboration

News

The CMS Collaboration at CERN is happy to announce the release of its fourth batch of open data to the public. With this release, which brings the volume of its open data to more than 2 PB (or two million GB), CMS has now provided open access to 100% of its research data recorded in proton-proton collisions in 2010, in line with the collaboration's data-release policy. The release also includes several new data and simulation samples. The new release builds upon and expands the scope of the successful use of CMS open data in research and in education.

(Image: Fermilab/CERN.)

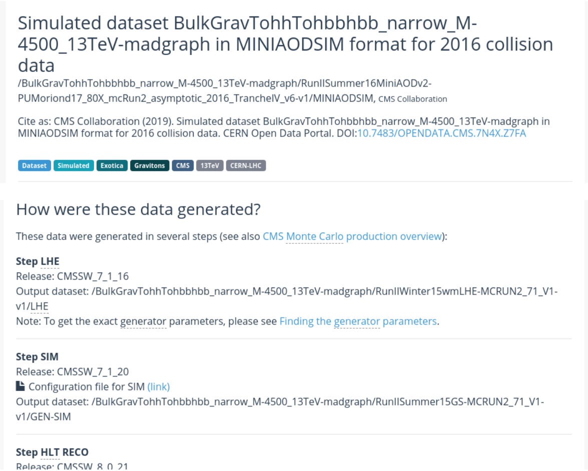


Release highlights

- This is the fourth release of high-level CMS open data, following release of around 50% of data from the LHC's Run 1: 2010 data in 2014, 2011 data in 2016, and 2012 data in 2017. This brings the volume of CMS open data to more than 2 PB.
- The release includes datasets prepared specifically for use in Machine Learning or in data science
  - A dataset derived from Run 2 simulation data is devoted to the challenge of event and object tagging in events with two b quarks produced from the decay of a Higgs boson. It is particularly difficult to distinguish this Higgs signal channel from the background.
  - Further datasets derived from Run 1 and Run 2 simulated data are devoted to identifying top quarks produced in events and to studying the flavour content of jets.
  - Another dataset is devoted to the challenge of particle tracking in the future era of high-luminosity collisions and is derived from simulations of collisions in the tracker after Phase 2 upgrades
- The parent datasets and production workflows for the ML samples also available for full reproducibility.
  - These include the first simulation samples in the "MiniAODSIM" format in use in Run 2 data analysis.
- Small samples of raw data are released, useful for testing of the data-processing chain and eventually reconstruction-algorithm development.
- Instructions are now available on how to generate simulated events in the open data environment.
- The release completes the 2010 data release with now all proton-proton data available publicly and adds some simulated data also for 2010 data taking.
- Contains datasets from early commissioning runs used in studies with CASTOR calorimeter and corresponding simulations.
- In addition to already available 2012 simulation data, large amount of 2012 simulation data of rarely used processes is now available on demand.
- Search functionality for simulation data is now available based on physics processes

Latest batch of CMS open data was released in Summer 2019

Example 1: Data provenance of simulated datasets



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

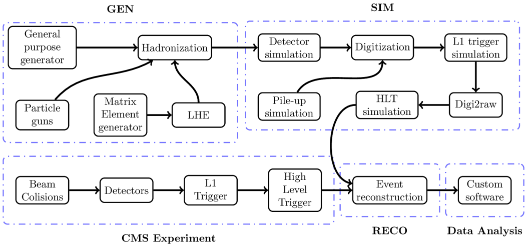
How were these data generated?

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

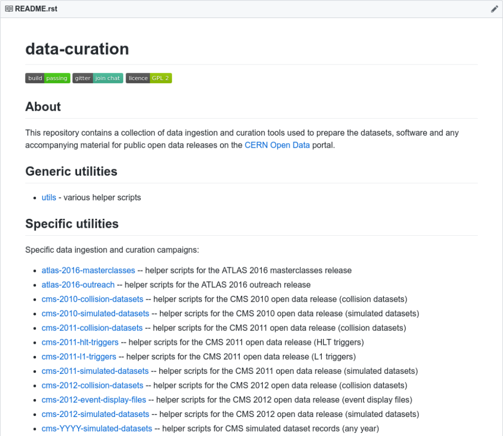
Step HLT RECO
Release: CMSSW_8_0_21



- 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

Capturing data provenance via ad-hoc curation techniques



README.rst

data-curation

build passing | gitter join chat | licence GPL 2

About

This repository contains a collection of data ingestion and curation tools used to prepare the datasets, software and any accompanying material for public open data releases on the CERN Open Data portal.

Generic utilities

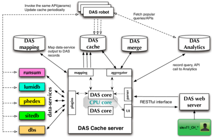
- utils - various helper scripts

Specific utilities

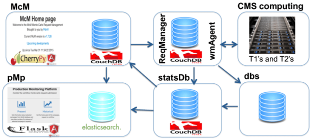
Specific data ingestion and curation campaigns:

- atlas-2016-masterclasses -- helper scripts for the ATLAS 2016 masterclasses release
- atlas-2016-outreach -- helper scripts for the ATLAS 2016 outreach release
- cms-2010-collision-datasets -- helper scripts for the CMS 2010 open data release (collision datasets)
- cms-2010-simulated-datasets -- helper scripts for the CMS 2010 open data release (simulated datasets)
- cms-2011-collision-datasets -- helper scripts for the CMS 2011 open data release (collision datasets)
- cms-2011-hlt-triggers -- helper scripts for the CMS 2011 open data release (HLT triggers)
- cms-2011-l1-triggers -- helper scripts for the CMS 2011 open data release (L1 triggers)
- cms-2011-simulated-datasets -- helper scripts for the CMS 2011 open data release (simulated datasets)
- cms-2012-collision-datasets -- helper scripts for the CMS 2012 open data release (collision datasets)
- cms-2012-event-display-files -- helper scripts for the CMS 2012 open data release (event display files)
- cms-2012-simulated-datasets -- helper scripts for the CMS 2012 open data release (simulated datasets)
- cms-YYYY-simulated-datasets -- helper scripts for CMS simulated dataset records (any year)

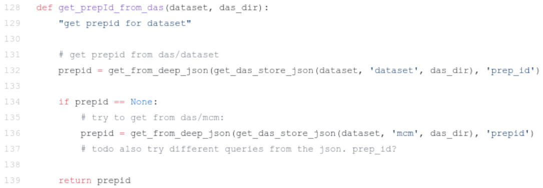
Dedicated data curation scripts



CMS DAS



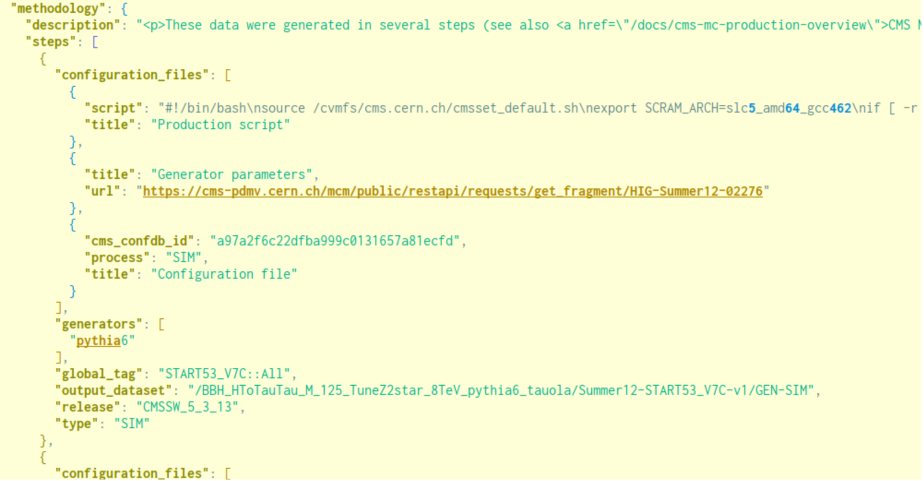
CMS McM



```
def get_prepId_from_das(dataset, das_dir):
    "get prepid for dataset"

    # get prepid from das/dataset
    prepid = get_from_deep_json(get_das_store_json(dataset, 'dataset', das_dir), 'prep_id')

    if prepid == None:
        # try to get from das/mcm:
        prepid = get_from_deep_json(get_das_store_json(dataset, 'mcm', das_dir), 'prepid')
        # todo also try different queries from the json. prep_id?

    return prepid
```

Mining several CMS collaboration sources

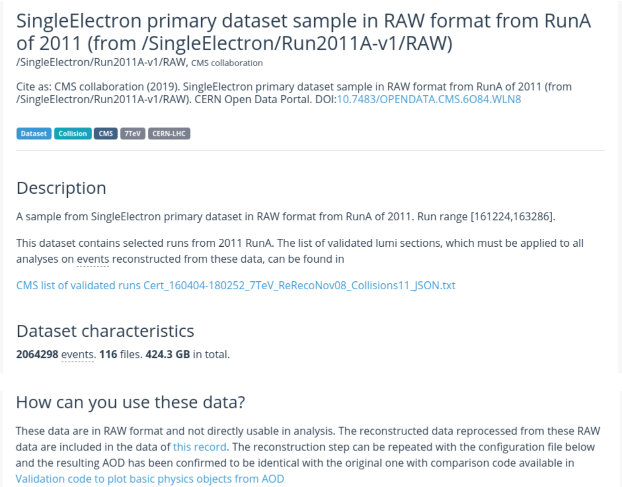
Harmonising year-dependent sources



```
"methodology": {
  "description": "<p>These data were generated in several steps (see also <a href=\"/docs/cms-mc-production-overview\">CMS
  "steps": [
    {
      "configuration_files": [
        {
          "script": "#!/bin/bash\nsource /cvmfs/cms.cern.ch/cmsset_default.sh\nexport SCRAM_ARCH=slc5_amd64_gcc462\nif [ -r
          "title": "Production script"
        },
        {
          "title": "Generator parameters",
          "url": "https://cms-pdmv.cern.ch/mcm/public/restapi/requests/get_fragment/HIG-Summer12-02276"
        },
        {
          "cms_confdb_id": "a97a2f6c22dfba999c0131657a81ecfd",
          "process": "SIM",
          "title": "Configuration file"
        }
      ],
      "generators": [
        "pythia6"
      ],
      "global_tag": "START53_V7C::All",
      "output_dataset": "/BBH_HToTauTau_M_125_TuneZ2star_8TeV_pythia6_tauola/Summer12-START53_V7C-v1/GEN-SIM",
      "release": "CMSSW_5_3_13",
      "type": "SIM"
    },
    {
      "configuration_files": [
```

From year-dependent DAS/McM information to year-independent Open Data JSON schema

Example 2: Raw data samples for 2010-2012 data



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

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

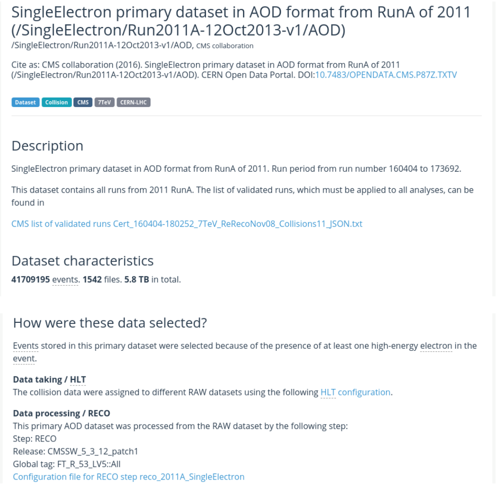
Dataset characteristics

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

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.

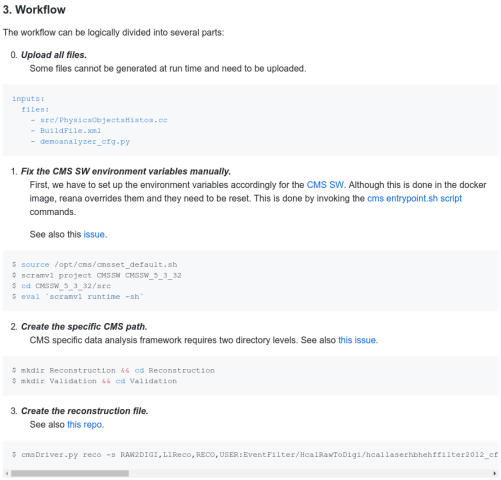
Data taking / HLT
The collision data were assigned to different RAW datasets using the following HLT configuration.

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

@tiborsimko

Can we reprocess raw data samples from 2010-2012?



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.

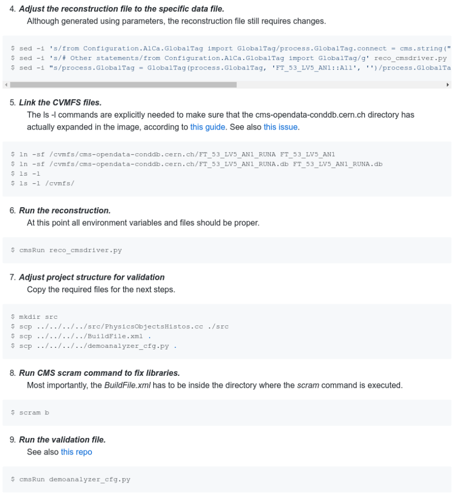
```
$ source /opt/cms/cmsset_default.sh
$ scramv1 project CMSSW CMSSW_5_3_32
$ cd CMSSW_5_3_32/src
$ eval `scramv1 runtime -sh`
```

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.

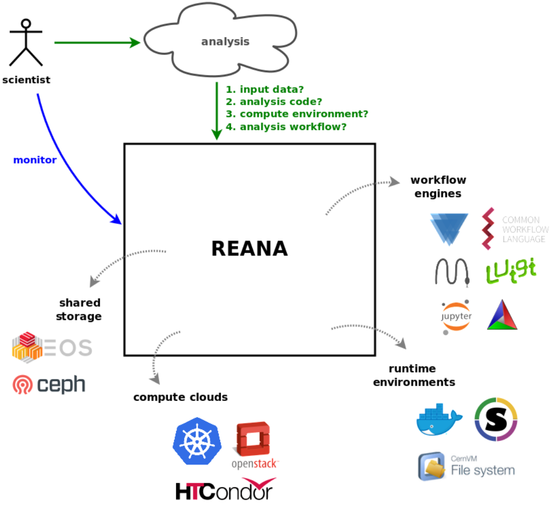
```
$ scram b
```

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

```
$ cmsRun demoanalyzer_cfg.py
```

Workflow steps to run CMS reconstruction in CMSSW environment

Running scientific workflows on containerised clouds

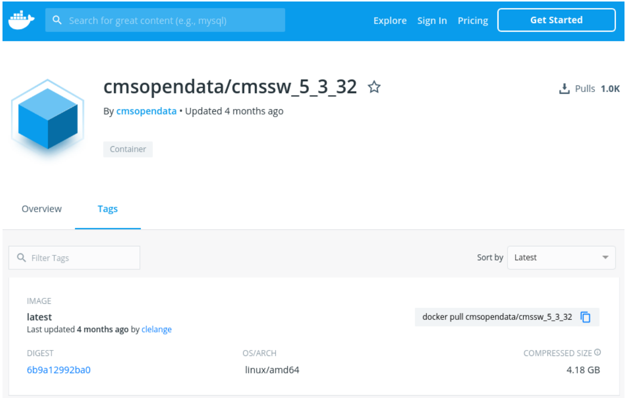


- REANA reproducible analysis platform `http://www.reana.io`
- multiple workflow systems (CWL, Serial, Yadage)
- multiple compute backends (Kubernetes, HTCondor, Slurm)
- multiple shared storage (Ceph, EOS, NFS)

reproducibility

code + data + environment + workflow

Preserving CMS software stack environment



cmsopendata/cmssw_5_3_32 ☆

Pulls 1.0K

By cmsopendata • Updated 4 months ago

Sort by Latest

IMAGE

latest

Last updated 4 months ago by clelange

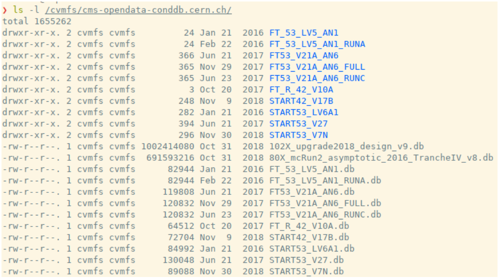
docker pull cmsopendata/cmssw_5_3_32

DIGEST: 6b9a12992ba0

OS/ARCH: linux/amd64

COMPRESSED SIZE: 4.18 GB

CMSSW docker image with "embedded" CVMFS



```
❯ ls -l /cvmfs/cms-opendata-conddb.cern.ch/
total 1655262
drwxr-xr-x. 2 cvmfs cvmfs        24 Jan 21  2016 FT_53_LV5_AN1
drwxr-xr-x. 2 cvmfs cvmfs        24 Feb 22  2016 FT_53_LV5_AN1_RUNA
drwxr-xr-x. 2 cvmfs cvmfs       366 Jun 21  2017 FT53_V21A_AN6
drwxr-xr-x. 2 cvmfs cvmfs       365 Nov 29  2017 FT53_V21A_AN6_FULL
drwxr-xr-x. 2 cvmfs cvmfs       365 Jun 23  2017 FT53_V21A_AN6_RUNC
drwxr-xr-x. 2 cvmfs cvmfs         3 Oct 20  2017 FT_R_42_V10A
drwxr-xr-x. 2 cvmfs cvmfs       248 Nov  9  2018 START42_V17B
drwxr-xr-x. 2 cvmfs cvmfs       282 Jan 21  2016 START53_LV6A1
drwxr-xr-x. 2 cvmfs cvmfs       394 Jun 21  2017 START53_V27
drwxr-xr-x. 2 cvmfs cvmfs       296 Nov 30  2018 START53_V7N
-rw-r--r--. 1 cvmfs cvmfs 1002414080 Oct 31  2018 102X_upgrade2018_design_v9.db
-rw-r--r--. 1 cvmfs cvmfs  691593216 Oct 31  2018 80X_mcRun2_asymptotic_2016_TrancheIV_v8.db
-rw-r--r--. 1 cvmfs cvmfs      82944 Jan 21  2016 FT_53_LV5_AN1.db
-rw-r--r--. 1 cvmfs cvmfs      82944 Feb 22  2016 FT_53_LV5_AN1_RUNA.db
-rw-r--r--. 1 cvmfs cvmfs     119808 Jun 21  2017 FT53_V21A_AN6.db
-rw-r--r--. 1 cvmfs cvmfs     120832 Nov 29  2017 FT53_V21A_AN6_FULL.db
-rw-r--r--. 1 cvmfs cvmfs     120832 Jun 23  2017 FT53_V21A_AN6_RUNC.db
-rw-r--r--. 1 cvmfs cvmfs      64512 Oct 20  2017 FT_R_42_V10A.db
-rw-r--r--. 1 cvmfs cvmfs      72704 Nov  9  2018 START42_V17B.db
-rw-r--r--. 1 cvmfs cvmfs      84992 Jan 21  2016 START53_LV6A1.db
-rw-r--r--. 1 cvmfs cvmfs     130048 Jun 21  2017 START53_V27.db
-rw-r--r--. 1 cvmfs cvmfs      89088 Nov 30  2018 START53_V7N.db
```

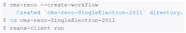
Condition data for open data analyses are available on "live" CVMFS

Automated reconstruction workflows

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

1 input parameters

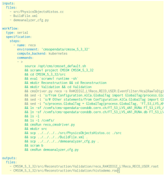
↓



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

2 workflow factory

→



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

workflow:
  type: serial
  specification:
    steps:
      - name: reco
        environment: 'cmsopendata/cmssw_5_3_32'
        compute_backend: kubernetes
        commands:
        - >
            source /opt/cms/cmsset_default.sh
            && scramv1 project CMSSW CMSSW_5_3_32
            && cd CMSSW_5_3_32/src
            && eval `scramv1 runtime -sh`
            && mkdir Reconstruction && cd Reconstruction
            && mkdir Validation && cd Validation
            && cmsDriver.py reco -s RAW2DIGI,L1Reco,RECO,USER:EventFilter/HcalRawToDigi
            && sed -i 's/from Configuration.AlCa.GlobalTag import GlobalTag/process.Glo
            && sed -i 's/# Other statements/from Configuration.AlCa.GlobalTag import Gl
            && sed -i "s/process.GlobalTag = GlobalTag(process.GlobalTag, 'FT_53_LV5_AN
            && ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA FT_53_LV5_A
            && ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA.db FT_53_LV
            && ls -l
            && ls -l /cvmfs/
            && cmsRun reco_cmsdriver.py
            && mkdir src
            && scp ../../../../src/PhysicsObjectsHistos.cc ./src
            && scp ../../../../BuildFile.xml .
            && scp ../../../../demoanalyzer_cfg.py .
            && scram b
            && cmsRun demoanalyzer_cfg.py

outputs:
  files:
    - CMSSW_5_3_32/src/Reconstruction/Validation/reco_RAW2DIGI_L1Reco_RECO_USER.root
    - CMSSW_5_3_32/src/Reconstruction/Validation/histodemo.root
```

3 reana.yaml

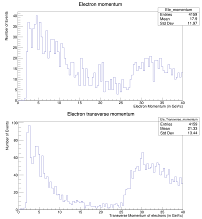
→

5 serving open data files

↓

4 run by REANA platform

→



6 output histograms

Parametrised workflow runnable on REANA reproducible analysis platform

Conclusions

CMS open data now contains detailed provenance information

- knowing "how the data came about" enhances current knowledge and future reuse
- capturing data provenance requires non-trivial information hunt and harmonisation
- *a posteriori* approach: running after $\sim$5 year old data and procedures
- *a priori* approach: ultra legacy run to generate preservation-friendly assets?

Successful RAW to AOD reconstruction tests on open data

- AOD reconstruction and histogram verification permitted to validate approach
- using non-production compute environment ensures reproducibility

CMS

`http://opendata.cern.ch`

opendata