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

Tibor Šimko[1,*], *Heitor Pascoal* de Bittencourt[2], *Edgar* Carrera[2], *Diyaselis* Delgado Lopez[2], *Clemens* Lange[2], *Kati* Lassila-Perini[2], *Adelina* Lintuluoto[2], *Lara* Lloret[2], *Tom* McCauley[2], *Jan* Okraska[1], *Daniel* Prelipcean[1], and *Mantas* Savaniakas[2]

[1]CERN Open Data team, CERN, Geneva, Switzerland
[2]CMS Collaboration

Abstract. In this paper we present the latest CMS open data release published on the CERN Open Data portal. The samples of raw datasets, collision and simulated datasets were released together with the detailed information about the data provenance. The data production chain covers the necessary compute environments, the configuration files and the computational procedures used in each data production step. We describe data curation techniques used to obtain and publish the data provenance information and we study the possibility to reproduce parts of the released data using the publicly available information. The present work demonstrates the usefulness of releasing selected samples of raw and primary data in order to fully ensure the completeness of information about data production chain for the attention of general data scientists and other non-specialists interested in using particle physics data for education or research purposes.

1 Introduction

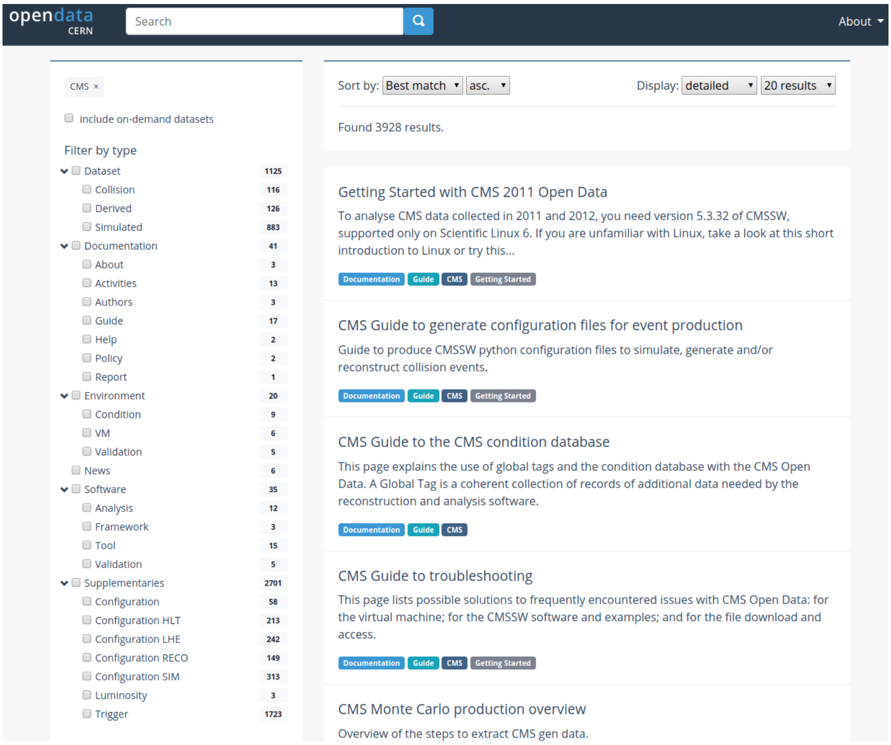
The CERN Open Data portal disseminates over two petabytes of data from particle physics [1]. It contains data from the four LHC collaborations based on their collaboration policies. The data is usuall y released to the public after a certain embargo period that serves to verify data quality. The released data is used for both education and research purposes [2].

The CMS experiment releases a large variety of open data on the portal. The data consists of collision and simulated datasets, the simplified derived datasets and event display files, the accompanying documentation, the virtual machines and software tools and analysis examples allowing to expore the data, and further supplementary material. The variety of data can be seen in Figure 1.

The open data releases are accompanied with rigorous data curation processes to provide enough documentation fro mthe data producers to the data consumers so that non-specialists, data scientists are able to understand and use the data. This is why capturing data provenance information, i.e. how to the data came to be, is crucial.

This paper describes procedures how the data provenance information was extracted and how it can be used for data validation and for facilitating future data reuse.

[*] e-mail: tibor.simko@cern.ch



CMS ×

□ include on-demand datasets

Filter by type

□ Dataset	1125
□ Collision	116
□ Derived	126
□ Simulated	883
□ Documentation	41
□ About	3
□ Activities	13
□ Authors	3
□ Guide	17
□ Help	2
□ Policy	2
□ Report	1
□ Environment	20
□ Condition	9
□ VM	6
□ Validation	5
□ News	6
□ Software	35
□ Analysis	12
□ Framework	3
□ Tool	15
□ Validation	5
□ Supplementaries	2701
□ Configuration	58
□ Configuration HLT	213
□ Configuration LHE	242
□ Configuration RECO	149
□ Configuration SIM	313
□ Luminosity	3
□ Trigger	1723

Sort by: Best match ▾ asc. ▾ Display: detailed ▾ 20 results ▾

Found 3928 results.

Getting Started with CMS 2011 Open Data

To analyse CMS data collected in 2011 and 2012, you need version 5.3.32 of CMSSW, supported only on Scientific Linux 6. If you are unfamiliar with Linux, take a look at this short introduction to Linux or try this...

Documentation Guide CMS Getting Started

CMS Guide to generate configuration files for event production

Guide to produce CMSSW python configuration files to simulate, generate and/or reconstruct collision events.

Documentation Guide CMS Getting Started

CMS Guide to the CMS condition database

This page explains the use of global tags and the condition database with the CMS Open Data. A Global Tag is a coherent collection of records of additional data needed by the reconstruction and analysis software.

Documentation Guide CMS

CMS Guide to troubleshooting

This page lists possible solutions to frequently encountered issues with CMS Open Data: for the virtual machine; for the CMSSW software and examples; and for the file download and access.

Documentation Guide CMS Getting Started

CMS Monte Carlo production overview

Overview of the steps to extract CMS gen data.

Figure 1. The richness of the CMS collections on the CERN Open Data portal. Note the search facets on the left indicating collision, derived and simulated datasets, various kinds of documentation, the computing environment and virtual machines, the software tools and example analyses, up to various kinds of supplementary material and configuration files.

2 Data provenance of simulated datasets

The CMS collaboration used several information systems to keep track of datasets. The two systems of particular interest are CMS DAS (Data Aggregation System) das and CMS McM (Monte Carlo) databases [4]. The systems store information about each dataset, its generation procedures, its parents, etc. For example, Figure 2 shows the steps involved in production of CMS simulated datasets.

The CMS DAS and McM systems store information that is being used in “live” physics analyses. The information is not meant to be understood by non-specialists and is somewhat ‘volatile’; for example the underlying data model can change from year to year, such that information about 2011 data found in one system may not be applicable to 2012 data that is hosted in other system. Releasing this information for non-specialists and general public therefore necessitates writing data curation scripts harvesting and harmonising this information.

We have developed custom curation scripts to perform metadata harvesting and harmonisation. For each released dataset, the scripts harvest available information from CMS DAS and CMS McM systems and store them into a common JSON schema model describing dataset provenance. Figure 3 provides one example showing the provenance information about the birth of the dataset. One can observed five different data generation steps, each indicating CMSSW software release version used, the Global Tag for condition database, the

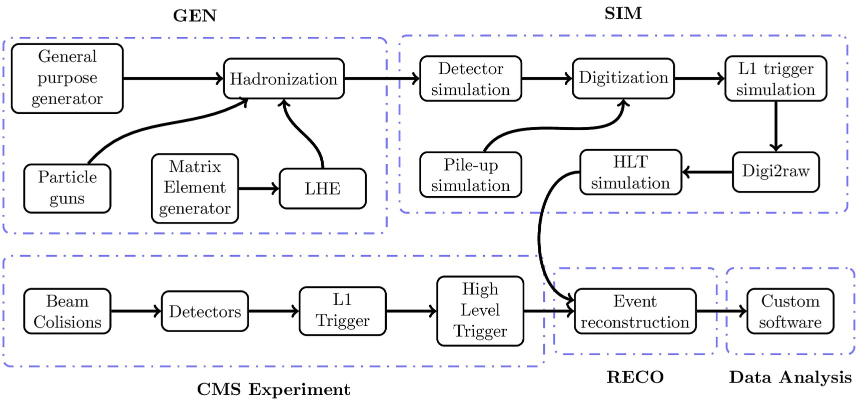


Figure 2. An overview of steps in production of CMS simulated datasets.

production script snippets or configuration files used, as well as the step output. The data provenance information constitutes a full recipe how the simulated data was generated which allows to understand it better at a glance.

3 Reprocessing raw data samples

The dataset provenance information obtained using procedures described in Section 2 constitute a "computational recipe" that allows to verify the correctness of released simulated data. The same principles apply not only to simulated data, but also to collision data. Here the data provenance chain allows to understand how the raw data taken by the CMS detector was later filtered for physics analyses.

The CMS open data releases contain samples of RAW data that allows to study this process. Figure 4 shows one example of released RAW data sample. Figure 5 shows corresponding AOD data that is used in physics analyses. Extracting dataset provenance information allows us to verify the reconstruction processes producing AOD data formats from RAW data samples.

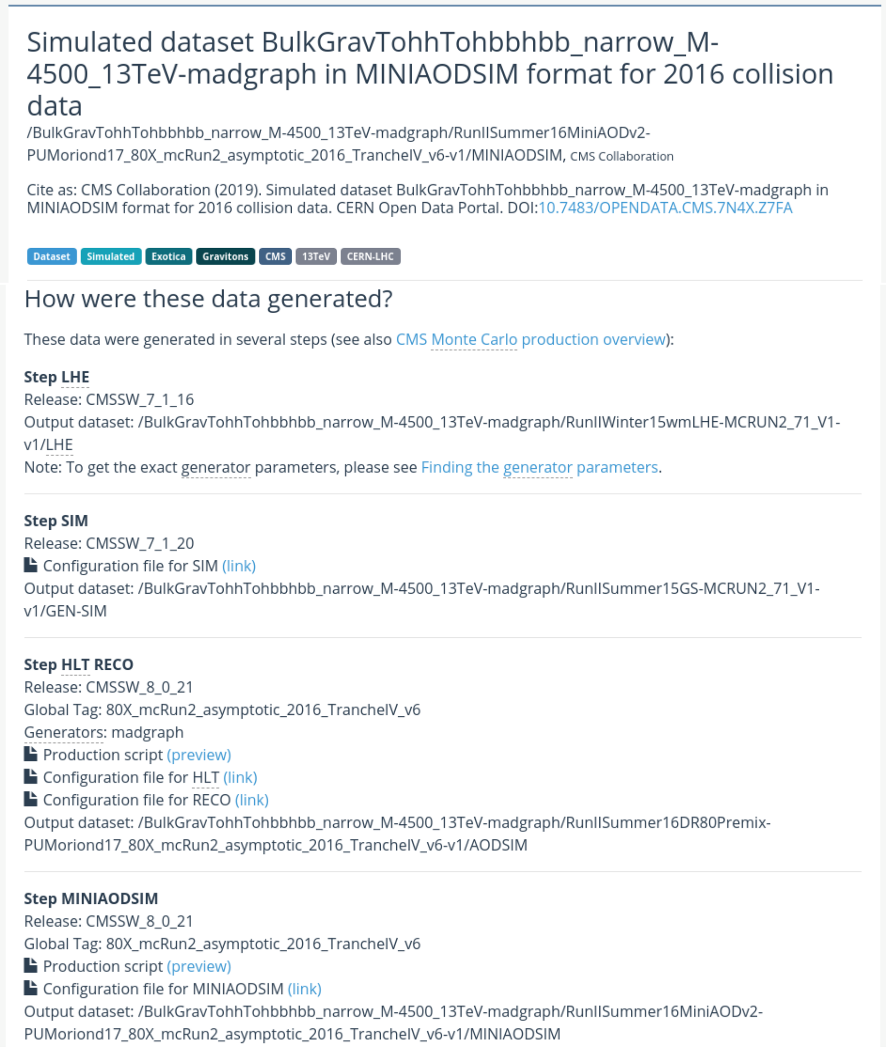
Figure 6 shows one detailed example of all the reconstruction workflow steps used to process RAW data to generad AOD data formats for physics analyses.

Let's run the same calculation on indepenent compute platform? Not easy. Need to capture database-like information, the condition database, which is stored on CVMFS. Solution: we developed CMSSW container images including CVMFS and we mount bigger condition database from CVMFS.

We have used REANA reproducible analysis platform [6] for which we have converted computational steps illustrated in Figure 6 into a structured workflow format. The example of produced histograms from this workflow is presented in Figure 7. We have compared thusly obtained histograms with released AOD files and have found a good match [7]. This allowed to validate obtained data provenance information using independent computing platform that was used to generate given data seven years ago.

4 Reconstruction workflow factory

The RAW sample processing to validate released AOD data formats is a recurrent need. The CMS collaboration released open data in yearly batches. The RAW reprocessing workflow



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

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
Global Tag: 80X_mcRun2_asymptotic_2016_TrancheIV_v6
Generators: madgraph
Production script (preview)
Configuration file for HLT (link)
Configuration file for RECO (link)
Output dataset: /BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer16DR80Premix-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/AODSIM

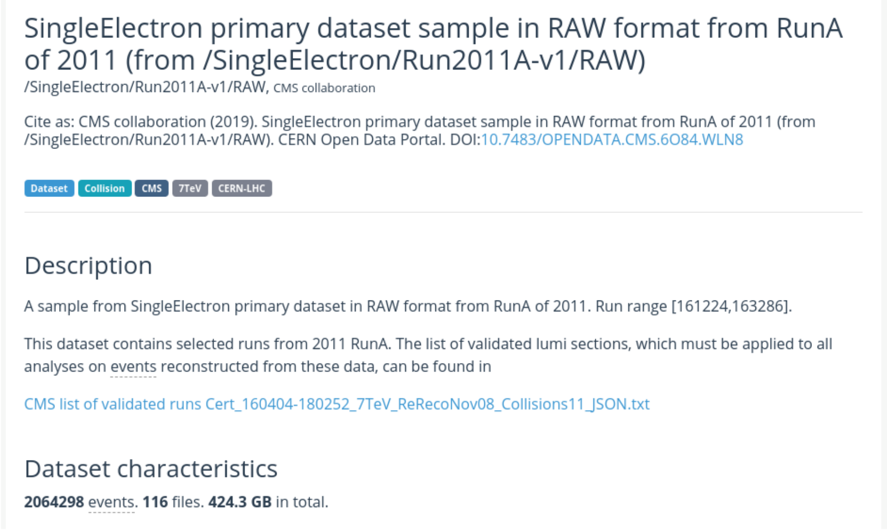
Step MINIAODSIM
Release: CMSSW_8_0_21
Global Tag: 80X_mcRun2_asymptotic_2016_TrancheIV_v6
Production script (preview)
Configuration file for MINIAODSIM (link)
Output dataset: /BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM

Figure 3. Example of a CMS simulated dataset available on the CERN Open Data portal with detailed provenance information displayed in the "How were these data generated?" section of the site.

can be thought of as taking two input variables: the data-taking year (e.g. 2010, 2011, 2012) and the dataset sample (e.g. Mu, SingleElectron, etc). We have therefore created a "workflow factory" that, given a wanted data-taking year and wanted dataset sample, generates the reconstruction workflow that can be run by the user. The command-line interaction goes like:

```
$ cms-reco --create-workflow --dataset DoubleElectron --year 20111
Created ‘cms-reco-DoubleElectron-2011‘ directory.
$ cd cms-reco-DoubleElectron-2011
$ reana-client run
```

The RAW-to-AOD data reconstruction workflow factory system is presented in Figure 8. It can be used to quickly generate validation workflows to verify the correctness of data provenance information about released open 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].

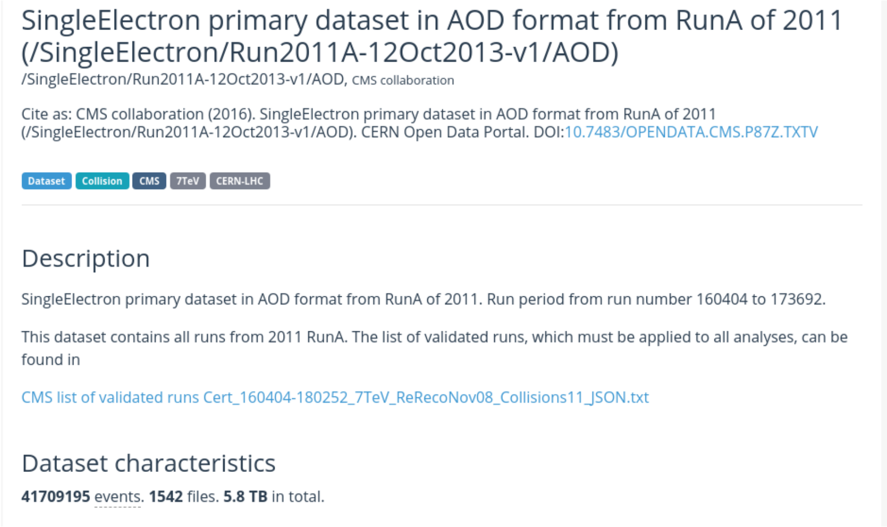
This dataset contains selected runs from 2011 RunA. The list of validated lumi sections, which must be applied to all analyses on events reconstructed from these data, can be found in

CMS list of validated runs Cert_160404-180252_7TeV_ReRecoNov08_Collisions11_JSON.txt

Dataset characteristics

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

Figure 4. Example of RAW sample available on the CERN Open Data portal. The corresponding reconstructed AOD dataset is visible in Figure 5.



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

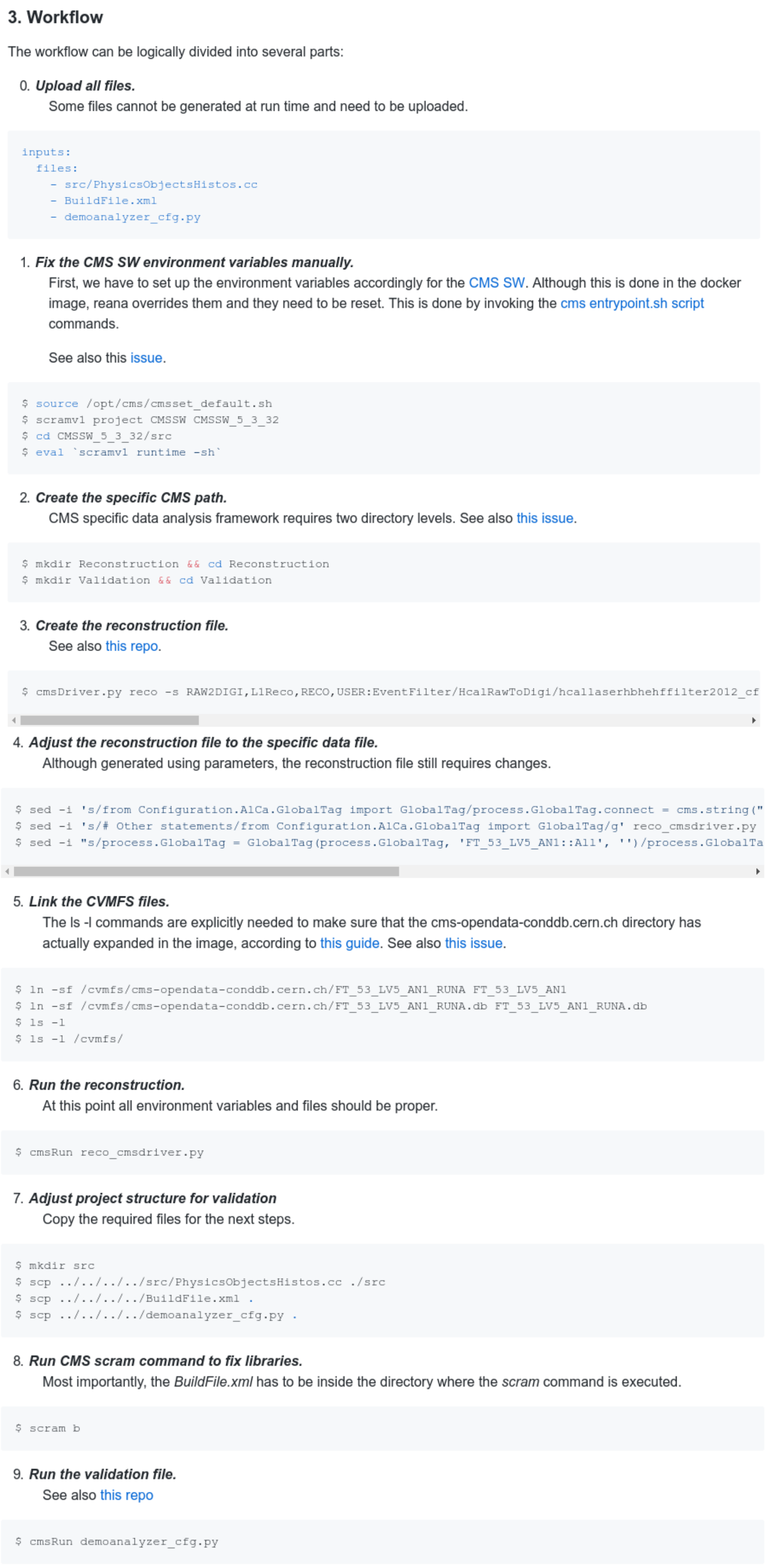
Dataset characteristics

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

Figure 5. Example of reconstructed AOD dataset released on the CERN Open Data portal. A part of this dataset is coming from the RAW data sample from Figure 4.

5 Conclusions

The CMS collaboration releases massive amounts of open data for research. This necessitates aggregating accompanying information about data and the context of data selection, validation and use. The capturing of data provenance information is crucial in understanding the data.



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
```

Figure 6. An individual reconstruction workflow and its runtime instructions.

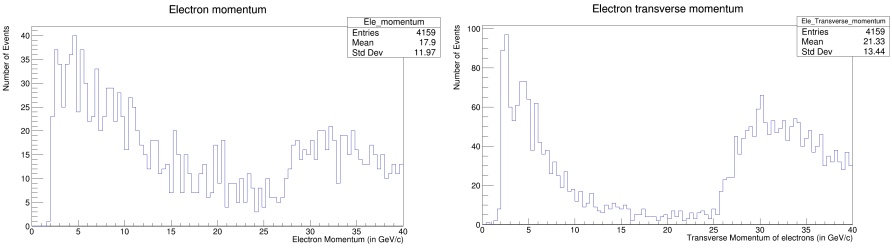
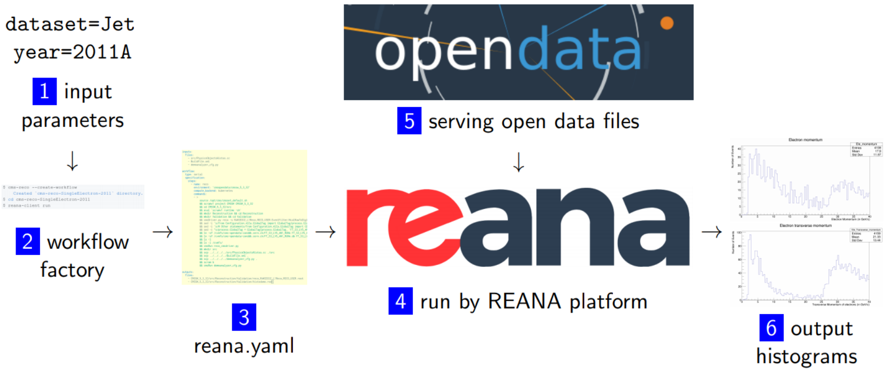


Figure 7. Example histograms produced by the reconstruction workflow.



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

```
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.Gl
          && 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_
          && ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA.db FT_53_L
          && 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
```

Figure 8. An overview of RAW-to-AOD data reconstruction workflow factory.

We have developed a set of curation scripts that mine CMS collaboration internal sources (DAS, McM) and aggregate the information in uniform JSON format for inclusion into the CERN Open Data portal. The released CMS data are accompanied with detailed information about provenance for most data-taking years. For some data-taking years, the information was not possible to extract due to changes in underlying CMS information sources. This highlights a need prepare future data reuse while the data-taking is active. The improvemens in this sense will be part of a future work.

We have also developed computational workflow factory for the REANA reproducible analysis platform that allow to verif the extracted dataset provenance information by running the data generation steps on indepedent containerised compute platform. We have shown an example of RAW to AOD process validation where we found a good match for data released many years ago. This demonstrates both the correctness of extracted data provenance information and the good reproducibility of data production workflows on independent computing platform.

References

[1] CERN Open Data Portal. `http://opendata.cern.ch`

[2] FIXME cite some Jessie Thaler et al papers on research use of open data

[3] FIXME cite CMS DAS

[4] FIXME cite CMS McM

[5] X. Chen, S. Dallmeier-Tiessen, R. Dasler, S. Feger, P. Fokianos, J. .B. Gonzalez, H. Hirvonsalo, D. Kousidis, A. Lavasa, S. Mele, D. Rodríguez, T. Šimko, T. Smith, A. Trisovic, A. Trzcinska, I. Tsanaktsidis, M. Zimmermann, K. Cranmer, L. Heinrich, G. Watts, M. Hildreth, L. Lloret Iglesias, K. Lassila-Perini, S. Neubert, “Open is not enough“, Nature Physics **15** 113–118 (2019). `https://www.nature.com/articles/s41567-018-0342-2`

[6] T. Šimko, L. Heinrich, H. Hirvonsalo, D. Kousidis, D. Rodríguez, “REANA: A system for reusable research data analyses”, EPJ Web of Conferences 214, 06034 (2019), `https://doi.org/10.1051/epjconf/201921406034`

[7] FIXME cite Diyaselis/Adelina/Kati poster on AOD validation