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

Tibor Šimko[1,*], *Heitor Pascoal* de Bittencourt[2], *Edgar* Carrera[3], *Diyaselis* Delgado Lopez[4], *Clemens* Lange[1], *Kati* Lassila-Perini[2], *Adelina* Lintuluoto[1,2], *Lara* Lloret Iglesias[5], *Thomas* McCauley[6], *Jan* Okraska[1], *Daniel* Prelipcean[1,7], and *Mantas* Savaniakas[8]

[1]CERN, Geneva, Switzerland
[2]Helsinki Institute of Physics, Finland
[3]Universidad San Francisco de Quito, Ecuador
[4]University of Puerto Rico, Puerto Rico
[5]Instituto de Física de Cantabria, Santander, Spain
[6]University of Notre Dame, US
[7]Technical University of Munich, Germany
[8]University of Vilnius, Lithuania

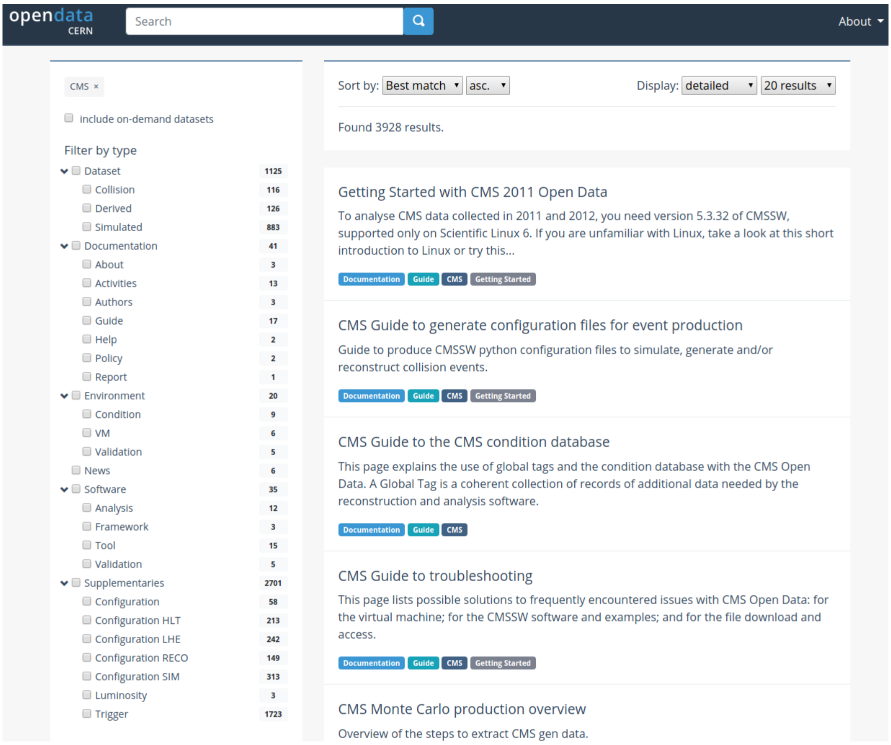
Abstract. In this paper we present the latest CMS open data release published on the CERN Open Data portal. Samples of collision and simulated datasets were released together with detailed information about the data provenance. The associated data production chains cover the necessary computing environments, 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 of reproducing 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 the 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 experiments [1]. It contains data from the four LHC collaborations based on their collaboration policies. The released data are used for both education and research purposes (for example [2], [3]) and are usually released to the public after a certain embargo period that allows for the exploitation of the data within the collaboration before the release and for the verification of data quality.

The CMS collaboration's policy on long-term data preservation, including the embargo terms, the reuse and open access policies is defined in [4]. The CMS policy was first approved by the CMS Collaboration Board in March 2012 and was updated in 2018 stating that, apart from the 50% of data that CMS will normally make available 3 years after data taking, 100%

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



□ 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.

of data will become available within 10 years. The CMS Collaboration Board can also, in exceptional circumstances, decide to release some particular data sets either earlier or later. The release latency has in practice been 5 years.

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

The open data releases are accompanied by rigorous data curation processes to provide enough documentation from the data producers to the data consumers so that the data can be understood and used by users external to the CMS collaboration. This paper describes the procedures by which the data provenance information was extracted and how it can be used for data validation and for facilitating future data reuse.

2 Data provenance of simulated datasets

The CMS collaboration uses several information systems to keep track of the datasets. The two systems of particular interest are CMS DAS (Data Aggregation System) [5] and CMS McM (Monte Carlo request management system) [6]. These databases store information

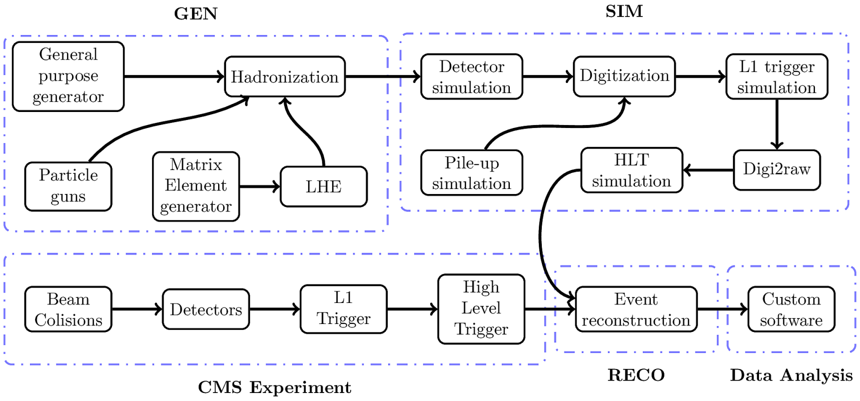


Figure 2. An overview of steps in production of CMS simulated datasets.

about each dataset, its generation procedures, its parents, etc. For example, Figure 2 shows the steps involved in production of CMS simulated datasets.

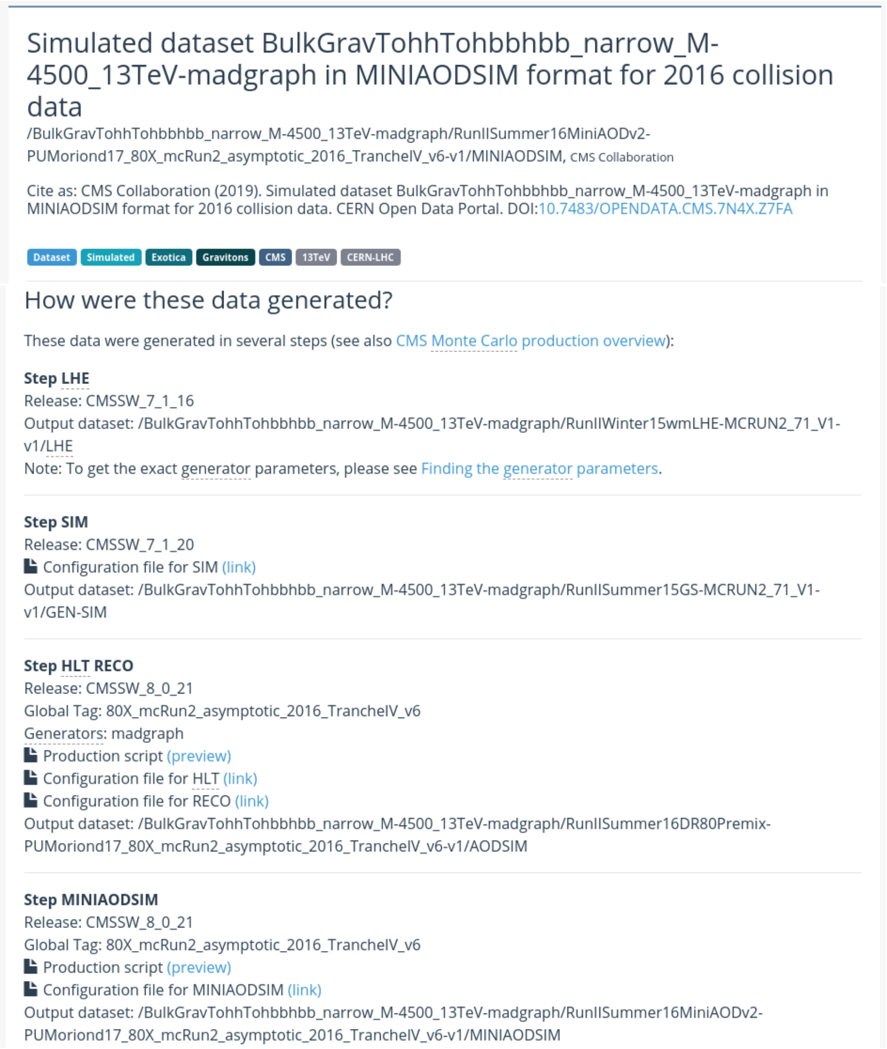
The CMS DAS and McM systems store information that is being used in “live” physics analyses. The information is not meant to be understood by non-specialists and is somewhat ‘volatile’; for example there have been changes in the CMS systems during years, such that procedures used to find information about 2010 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 the general public therefore necessitates writing data curation scripts that harvest and harmonise 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 combine them into a common JSON schema model describing dataset provenance. Figure 3 provides one example showing the provenance information about the production steps of a dataset. One can observe five different data generation steps, each indicating the CMS Offline Software (CMSSW) release version used, the Global Tag i.e. the additional conditions data, the production script snippets or configuration files used, as well as the output step. The data provenance information constitutes a full recipe on how the simulated data were generated thus providing the full history of these data.

3 Reprocessing raw data samples

The dataset provenance information obtained using the procedures described in Section 2 constitutes a “computational recipe” that allows for the replication of the processes used to generate released simulated data. The same principles apply not only to simulated data, but also to collision data. Here the data provenance chain allows us to understand how the raw data taken by the CMS detector were processed into a format appropriate for physics analyses.

The CMS open data releases contain samples of RAW data that allow us to study this process. Figure 4 shows one example of a released RAW data sample. Figure 5 shows corresponding Analysis Object Data (AOD) that are used in physics analyses. Extracting dataset provenance information allows us to repeat the reconstruction processes producing AOD data formats from RAW data samples.



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

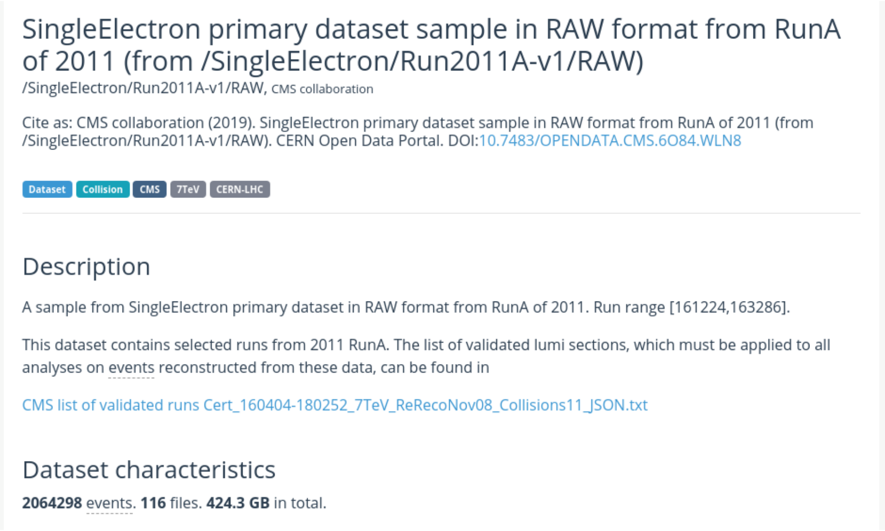
Step HLT RECO
Release: CMSSW_8_0_21
Global Tag: 80X_mcRun2_asymptotic_2016_TrancheIV_v6
Generators: madgraph
Production script (preview)
Configuration file for HLT (link)
Configuration file for RECO (link)
Output dataset: /BulkGravTohhTohbbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer16DR80Premix-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/AODSIM

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.

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

Running the same reconstruction workflow on an independent computing platform necessitates the replication of the original computing environment as best as possible. We have taken advantage of the container technology which allows us to encapsulate computing environments as Docker containers. The developed `cmsopendata/cmssw` container images [7] were published on Docker Hub and provide a complete CMSSW computing environment needed to work with CMS open data. Additional runtime data, such as the condition database, are stored independently in the CVMFS software distribution service from where it is read "live" during workflow execution.



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].

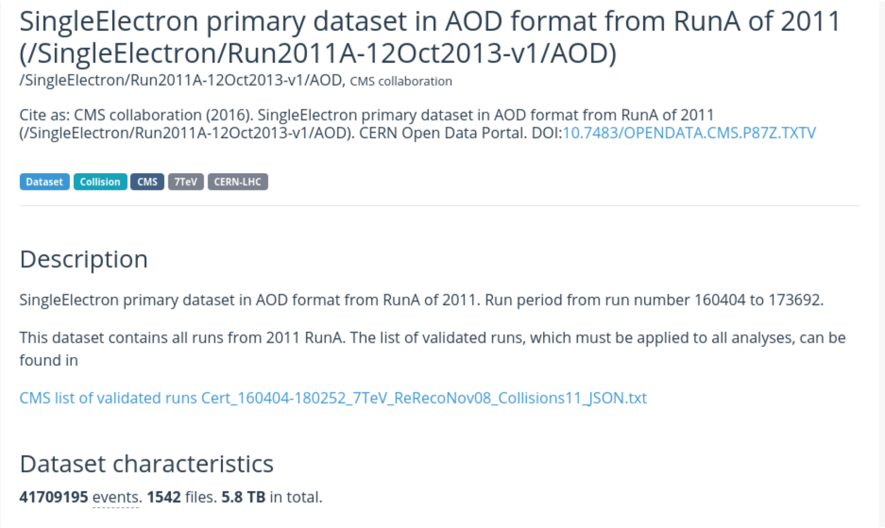
This dataset contains selected runs from 2011 RunA. The list of validated lumi sections, which must be applied to all analyses on events reconstructed from these data, can be found in

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 a RAW data sample available on the CERN Open Data portal. This sample corresponds with the reconstructed AOD dataset that is shown 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

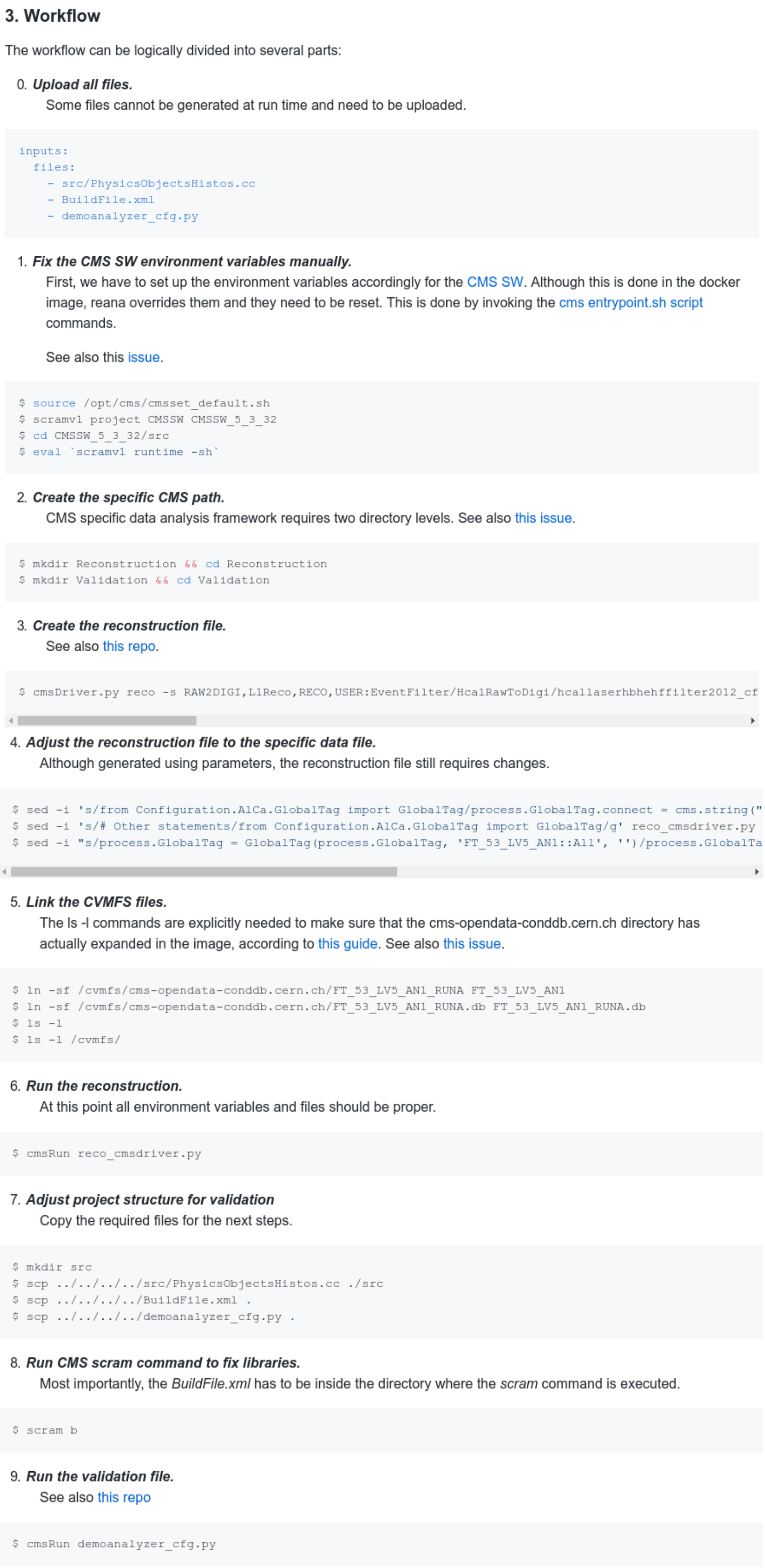
CMS list of validated runs Cert_160404-180252_7TeV_ReRecoNov08_Collisions11_JSON.txt

Dataset characteristics

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

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

We have used the REANA reproducible analysis platform [8] for which we have converted computational steps illustrated in Figure 6 into a structured workflow format. One example of histograms produced from this workflow is presented in Figure 7. We have compared histograms we obtained using this workflow with released AOD files and have found a good match. This allowed us to demonstrate that reprocessing of preserved LHC Run1 RAW data from 2010–2011 is possible and could be repeated in case of necessity.



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.

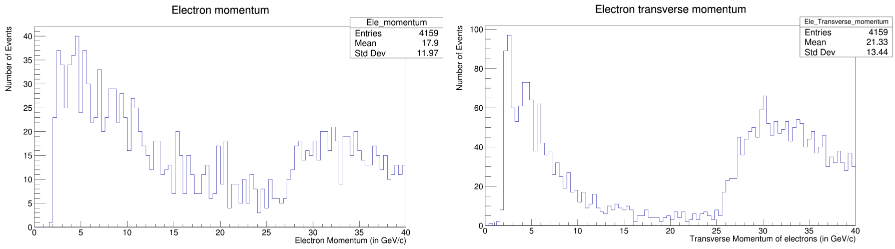
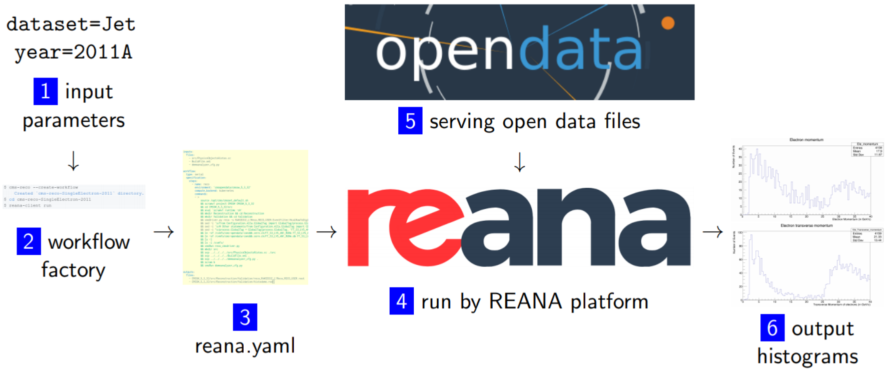


Figure 7. Example histograms produced by the reconstruction workflow.



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

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

4 Reconstruction workflow factory

The processing of RAW samples into the AOD data format can be used as a test workflow to validate the functionality of the preserved computing environment. Running such RAW sample processing test workflows is a recurrent need; the CMS collaboration releases open data in yearly batches. The RAW reprocessing workflow 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 desired data-taking year and a desired dataset sample, generates the reconstruction workflow that can be run by the user. An example of a command-line interaction is as follows:

```
$ cms-reco --create-workflow --dataset DoubleElectron --year 2011
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.

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 for understanding the data.

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

We have also developed a computational workflow factory for the REANA reproducible analysis platform that allows us to verify the extracted dataset provenance information by running the data generation steps on an independent 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 good reproducibility of data production workflows using an independent computing platform.

References

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

[2] Larkoski, Andrew, et al. "Exposing the QCD splitting function with CMS open data." Physical review letters 119.13 (2017): 132003.

[3] Tripathee, Aashish, et al. "Jet substructure studies with CMS open data." Physical Review D 96.7 (2017): 074003.

[4] CMS data preservation, reuse and open access policy (2018). CERN Open Data Portal, `DOI:10.7483/OPENDATA.CMS.7347.JDWH`

[5] Kuznetsov, V., Metson, S., & Evans, D. (2010). The CMS data aggregation system (No. CMS-CR-2010-036). `https://doi.org/10.1016/j.procs.2010.04.172`

[6] G. Boudoul et al, “Monte Carlo Production Management at CMS” J. Phys.: Conf. Ser. 664 072018 (2015)

[7] C. Lange, “cmsopendata/cmssw docker container images“, Docker Hub, 2019. `https://hub.docker.com/u/cmsopendata`

[8] 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`