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

Tibor Šimko¹ Heitor Pascoal de Bittencourt² Edgar Carrera² Clemens Lange² Kati Lassila-Perini² Lara Lloret² Tom McCauley² Jan Okraska¹ Daniel Prelipcean¹ Mantas Savaniakas²

on behalf of the CERN Open Data team and the CMS Collaboration

¹CERN Open Data team ²CMS 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

Developed by CERN-IT in close collaboration with Experiments

http://opendata.cern.ch
Education-oriented use cases

Interactive event display and histogramming for derived datasets
Research-oriented use cases

- Run CernVM Virtual Machines
- Run realistic physics analysis examples
Enables independent theoretical research

Over twenty papers citing CMS open data...that the CMS collaboration start to cite!

@tiborsimko
New CMS open data release

Latest batch of CMS open data was released in Summer 2019

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.
Example 1: Data provenance of simulated datasets

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

/CbgKerLiTohbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer16MiniAODv2.


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: /BulkGravTohhTohbhbb_narrow_M-4500_13TeV-madgraph/RunIIWinter15swmLHE-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: /BulkGravTohhTohbhbb_narrow_M-4500_13TeV-madgraph/RunIISummer15GS-MCRUN2_71_V1-V1/GEN-SIM

Step HLT RECO
- Release: CMSSW R 71

Data records come with full provenance information

- full capture of data generation steps
- full capture of compute environments
- full capture of configuration files
- full capture of production scripts
Capturing data provenance via ad-hoc curation techniques

Data-curation

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-simulated-datasets — helper scripts for the CMS 2011 open data release (simulated 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-2013-simulated-datasets — helper scripts for CMS simulated dataset records (any year)

Dedicated data curation scripts

Mining several CMS collaboration sources

@tiborsimko
Harmonising year-dependent sources

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)


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 this data, can be found in

CMS list of validated runs Cert_160404-180252_7TeV_ReRecoNov08_Collisions11.JSON.txt

Dataset characteristics

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

AOD
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.

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_9_3_32
   $ cd CMSSW_9_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 --> RAWDIGI,L1Reco,RECO,USER:EventFilter/HcalRawToDigit/HcalIsolatedHBHEfilter2012_cfi
   ```

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

   ```
   $ sed -i '/Other statements/Configuration.AiCa.GlobalTag import GlobalTag/g' reco_cmsdriver.py
   $ sed -i 's/Configuration.AiCa.GlobalTag = globalTag(process.GlobTag), "TT_T2_JSON_HLT/ d"
   ```

5. Link the CMSSW 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 -s /afs/cern.ch/cms-opendata-conddb.cern.ch/TT_T2_JSON_HLT/ Run1A/ RUN1A
   $ ln -s /afs/cern.ch/cms-opendata-conddb.cern.ch/TT_T2_JSON_HLT/ Run1B/ RUN1B
   $ ls -l
   ```

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
   $ cp .../.../.../PhysicsObjectsundance_cxx ./src
   $ cp .../.../.../demoAnalyzer_cff.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_cff.py
   ```
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

CMSSW docker image with “embedded” CVMFS

Condition data for open data analyses are available on “live” CVMFS
Automated reconstruction workflows

1 input parameters

2 workflow factory

3 reana.yaml

4 run by REANA platform

5 serving open data files

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 ∼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

http://opendata.cern.ch