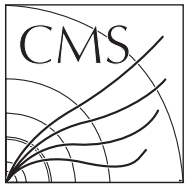
Available on CMS information server

CMS CR -2015/056



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

11 May 2015

Open access to high-level data and analysis tools in the CMS experiment at the LHC

A Calderon, D Colling, A Huffman, K Lassila-Perini, T McCauley, A Rao, A Rodriguez-Marrero and E Sexton-Kennedy for the CMS Collaboration

Abstract

The CMS experiment, in recognition of its commitment to datapreservation and open access as well as to education and outreach, hasmade its first public release of high-level data up to half of theproton-proton collision data at 7 TeV from 2010 in CMS Analysis ObjectData format. CMS has prepared, in collaboration with CERN and theother LHC experiments, an open data web portal based on Invenio. Theportal provides access to CMS public data as well as to analysis toolsand documentation for the public. The tools include an event displayand histogram application that run in the browser. In addition avirtual machine is available which contains a CMS software environmentalong with XRootD access to the data. Within the virtual machine thepublic can analyse CMS data; example code is provided. We describe theaccompanying tools and documentation and discuss the first experienceof data use.

Presented at *CHEP2015 21st International Conference on Computing in High Energy and Nuclear Physics*

Open access to high-level data and analysis tools in the CMS experiment at the LHC

A Calderon[1], D Colling[2], A Huffman[2], K Lassila-Perini[3], T McCauley[4], A Rao[5], A Rodriguez-Marrero[1] and E Sexton-Kennedy[6]

[1] IFCA, CSIC-Univ. de Cantabria, Santander, Spain
[2] Department of Physics, Imperial College, London, UK
[3] Helsinki Institute of Physics, Helsinki, Finland
[4] University of Notre Dame, Notre Dame, IN, USA
[5] University of the West of England, Bristol, UK
[6] Fermi National Accelerator Laboratory, Batavia, IL, USA

E-mail: k.lassila-perini@cern.ch

Abstract. The CMS experiment, in recognition of its commitment to data preservation and open access as well as to education and outreach, has made its first public release of high-level data under the CC0 waiver: up to half of the proton-proton collision data (by volume) at 7 TeV from 2010 in CMS Analysis Object Data format. CMS has prepared, in collaboration with CERN and the other LHC experiments, an open-data web portal based on Invenio. The portal provides access to CMS public data as well as to analysis tools and documentation for the public. The tools include an event display and histogram application that run in the browser. In addition a virtual machine containing a CMS software environment along with XRootD access to the data is available. Within the virtual machine the public can analyse CMS data; example code is provided. We describe the accompanying tools and documentation and discuss the first experiences of data use.

1. Introduction

The Compact Muon Solenoid or CMS [1] is one of two general-purpose experiments at the Large Hadron Collider (LHC) at CERN. Since 2010 CMS has collected around 28 fb^{-1} of proton-proton collision data at center-of-mass energies up to 8 TeV as well as data from proton-lead and lead-lead collisions. Analysis of these data have produced almost 400 published papers describing searches for new physics phenomena, measurements of known processes, as well as discovery of the Higgs boson [2].

In recognition of the importance of data preservation, re-use, and open access, CMS has approved a policy (found via [3]) that defines the collaboration's approach. The policy covers many levels, from a commitment to publication in open-access journals, to release of reconstructed data, to the preservation and release of software and documentation needed for reconstruction and analysis. In accordance with this policy, CMS has released, simultaneously with the release of the CERN Open Data Portal [4] in November 2014, a first large set of reconstructed data for public use. This dataset is around 27 TB of 2010 proton-proton collision data at 7 TeV.

2. CMS open data release

The newly released data are in CMS Analysis Object Data (AOD) format [5], which is the format used by CMS physicists for most analyses of CMS data in Run 1 (2010-2012). The data were collected during 2010's "RunB", the latter of the two data-taking periods of the year, from mid-September to the end of October 2010, and corresponding to an integrated luminosity of approximately 38 pb^{-1}. These high-level data are divided into so-called primary datasets, to which the data are directed based on the online event selection results. The dataset names, such as MinimumBias, Mu, Electron, and MultiJet (to give a few examples), reflect the selection criteria.

The released data were reconstructed in 2011 with CMS software (CMSSW) [6] version 4.2 running on Scientific Linux CERN 5 [7]. This is the latest complete reconstruction of the data taken in 2010. The data were not specifically prepared for the public release. Therefore, acknowledging the difficulty of using these data and in order to ease further re-use, CMS has provided, together with the analysis software in a virtual machine (VM), a basic set of instructions on the usage of the data as well as the necessary condition database needed by the software, along with a list of validated runs.

While the release of the corresponding simulated Monte Carlo datasets is foreseen for future releases of CMS data, it was not done for this first release, because no complete set of simulated data exists for this version of CMSSW (4.2). The simulated datasets for 2010 were only available for earlier CMSSW releases and their release as such was not considered useful. For the data taken in 2011 and 2012, part of which are to be made publicly available at the next data release, the simulated datasets are available with the same CMSSW release as the collision data, and will be released together with the data.

In addition to the primary datasets, CMS provides some examples of further reprocessed data derived from the primary datasets. These derived datasets are meant to be used with CMS analysis software (where reprocessing reduced the time needed for the final analysis) or with online web applications (where reprocessing reduced the complexity of the data in terms of content and format). A reduction of the Mu [8] and Electron [9] primary datasets is provided as derived data [10, 11]. These datasets contain only electron and muon candidates respectively, which are extracted from the AOD files with standard CMS software tools, and they can be analysed with an example code available with the open data release.

The derived datasets include an earlier release of a small, selected amount of data for use in education and outreach. These datasets were reduced to the level of four-vectors and contain $\text{J}/\psi, \Upsilon, \text{W}$, and Z candidates as well as general two-muon and two-electron events. In order to make usage as straight-forward and simple as possible the data were released in human-readable, text-based formats such as CSV (comma-separated variable) and JSON (JavaScript Object Notation) and have formed the core of the successful International Physics Masterclass program aimed at high-school students around the world [12].

All of the open data provided by CMS are released under the CC0 waiver [13], thus liberating the data into the public domain. Each dataset stored on the CERN Open Data Portal, described below, are also assigned Digital Object Identifiers (DOIs), which make the datasets citable objects (like in this very paper!).

3. CERN Open Data Portal

CERN, in collaboration with CMS and the other LHC experiments, has prepared an open data portal [4] from which the data and tools for the public are available. The portal itself is built using CERN's Invenio digital library software [14] as its base. Invenio provides document organization, search capability, and handling of metadata. CMS relies on CERN support and services provided through the portal for legacy data storage, access to and distribution of the data, and security and bandwidth restrictions for public access.

The portal builds upon the previous successes of public release of data for education and outreach but goes further by including the possibility for more in-depth, complex analyses, with the high-level data now made public by CMS. The portal is therefore divided into two sections – "Education" and "Research" – to which the material is assigned depending on its potential use and "degree of difficulty". The main entry view is shown in Figure 1. This division is designed to ease the access to material of interest depending on the user's intended usage. The portal directs the user to simplified web applications on the "Education" side and to detailed instructions on how to get started with analyses using CMS data on the "Research" side. The material is also ordered into different collections. Contributions from the four major LHC experiments – ALICE, ATLAS, CMS, and LHCb – are included.

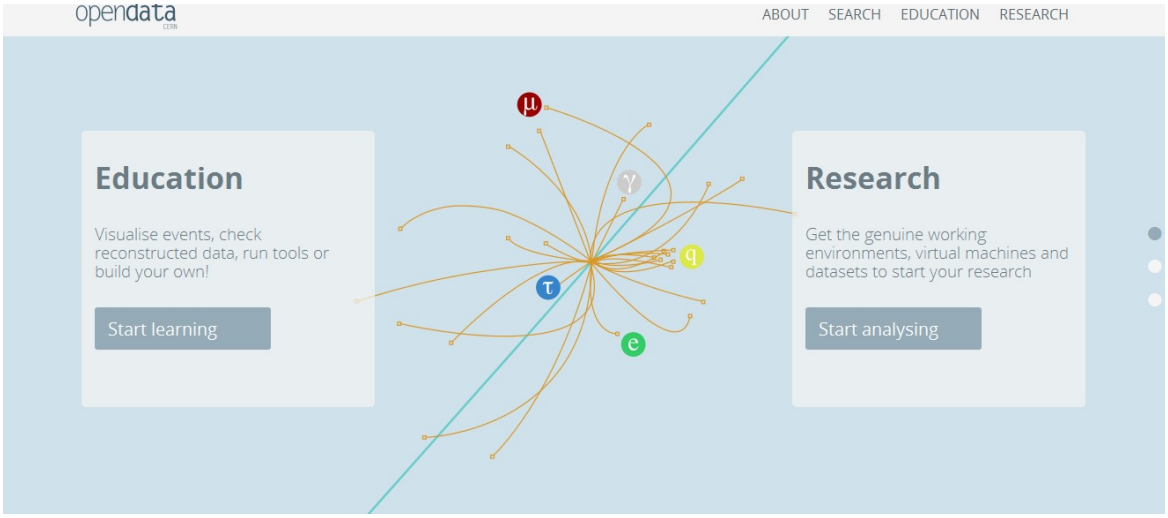


Figure 1. The main entry to the CERN Open Data Portal.

For CMS the main collections are currently:

- CMS Primary Datasets
- CMS Derived Datasets
- CMS Tools
- CMS Open Data Instructions
- CMS Learning resources.

The two first collections contain the data as described in the previous section. Each primary dataset record includes metadata fields including information of the provenance of the data and instructions for its usage. Records in the data collections may point to a record in the tools collections containing the software, which either can be used to study the data (in case of primary datasets) or has been used to generate the data (in case of derived datasets).

The software records in the tools collections can also act as examples, from which external users can build and extend their own applications. The open-source examples include code for implementing a simple selection of dimuons from the Mu sample [8] and generating the output of the four-vector information to a CSV file [15], and code for conversion of events into the format suitable for viewing in the event display described in the next section. Also included is code for doing a simple analysis of Z decays to two leptons and ZZ decays to four leptons [16]; the best

Z candidates are selected by performing different quality selections on the leptons (muons and electrons) and pairing those of high transverse momentum and opposite charge.

The tools collection also provides the source code for different web applications, described in the following section, that run on the derived datasets available from the portal, and the downloadable VM image built on CernVM [17] containing the CMS software and computing environment with which the primary datasets can be accessed without direct download (served via XRootD [18]).

The instructions collection pulls together the different tutorial and instruction pages of the portal. The learning resources collection links to various external resources that already use CMS open data and are readily usable in an educational context (teaching or learning).

4. Web applications

Two applications using the open data are available for use within the portal: an event display and a histogram application. Both are client-side applications running in the browser.

4.1. Event display

The event display [19] is an adaptation of an application [20] used by the public in educational programs such as the International Physics Masterclasses [12]. This particular version of the display is written in JavaScript, HTML, and CSS. The 3D rendering is canvas-based and is done using the pre3d library [21]. A screenshot of an event in the display can be seen in Figure 2.

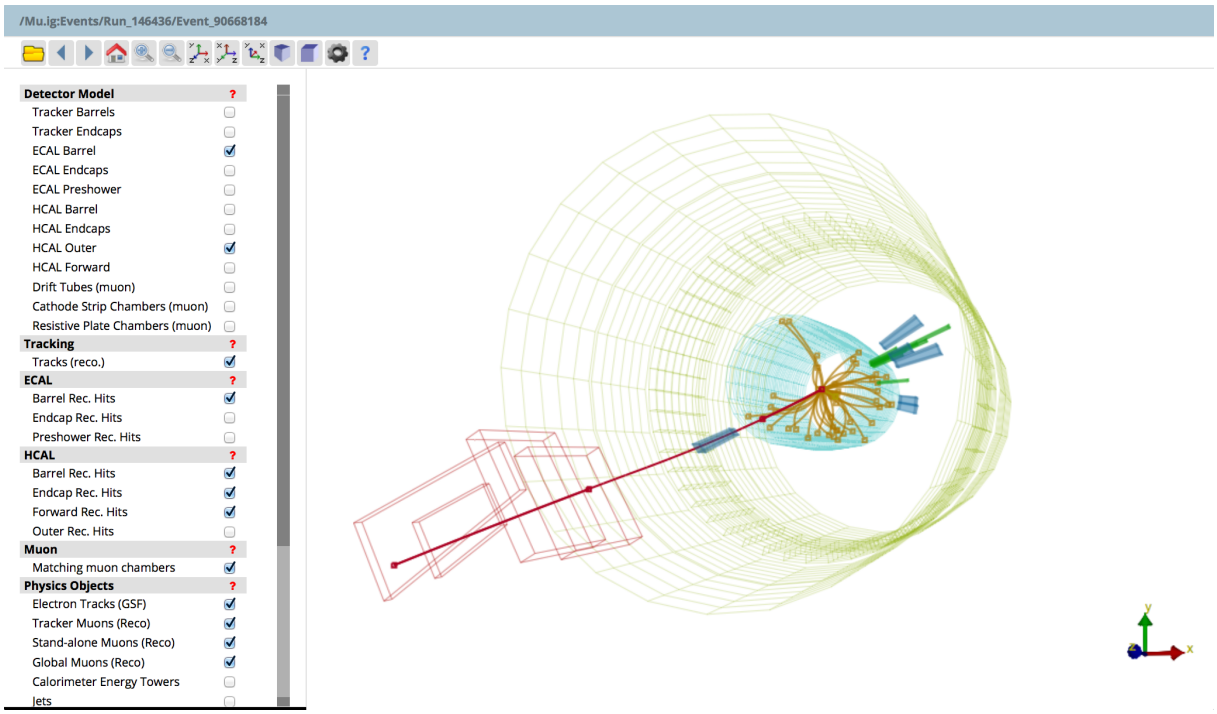


Figure 2. A screenshot of an event display from the Mu primary dataset.

The input file format for the display is the ig format [22], which is a zip archive containing a JSON file for each event. In Invenio, a record that contains a file in this format can be viewed in the browser with the invenio-previewer-ispy [23]. In the event display all of the ig files available in the portal as records are available for viewing. There are fourteen CMS primary datasets currently available on the portal; twenty-five events from each of the primary datasets

are available in ig format for viewing. In addition to example events from the primary datasets, additional events from previously-released datasets are available: these include di-lepton and W events used in the masterclasses.

Since the release of the portal, a new version of the display has been developed [24] that uses bootstrap.js for the user interface and WebGL for 3D rendering (defaulting to canvas-based rendering when WebGL is not available). This version will eventually replace the canvas-based version currently available.

4.2. Histogram tool

A subset of events from several of the open datasets are available in CSV format, where each line contains event information at the level of four-vectors.

The information contained in these CSV files is available for view in an interactive histogram application [25]. The application uses d3 [26] for creation of the histogram information and the jQuery-based library Flot [27] for the graphics. An example of histograms for events where a W boson decays into an electron and a neutrino can be seen in Figure 3. Current functionality includes logarithmic axes, range selection along the horizontal axis, and the ability to change the bin width. In the future one will be able to, for multiple plots, make correlated selection of parameter ranges.

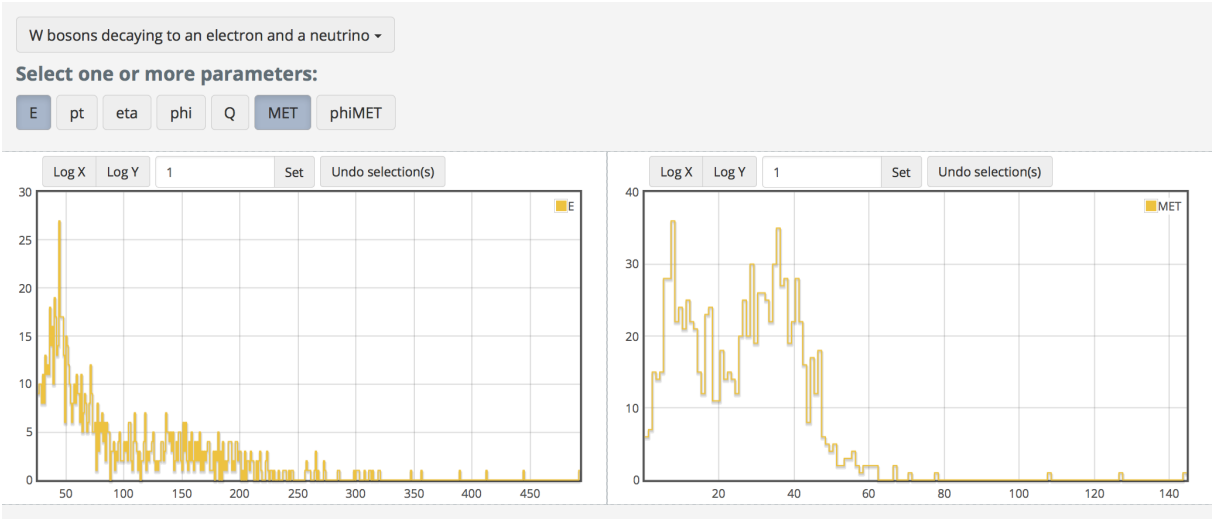


Figure 3. A screenshot of the histogram tool in the open data portal.

5. Communication and outreach

Instead of adopting the "build it and they will come" approach, it was decided from the start that appropriate means of dissemmination would need to be exercised in order to inform interested parties about the existence of the CERN Open Data Portal with CMS open data, and engage with said parties across the globe. Various audiences were considered, including the press, educators (school and university), physicists, as well as members of the general public. Accordingly, communications and outreach efforts were handled as described below.

5.1. Formal communications: CERN press release

A press release was drafted by all involved partners and issued by CERN in November 2014 [28] to announce the launch of the CERN Open Data Portal to the widest possible audience. CMS

also drafted an official statement [29], which was linked from the CERN press release, covering the collaboration's motivations for releasing the data and the expected audiences who would use them. The press release generated worldwide coverage, and the launch of the portal was covered in several media publications. CERN's official tweet [30] announcing the portal was the third-most re-tweeted post of 2014, which showed that there was interest among the wider community of science enthusiasts.

5.2. Engagement through social media

To leverage this wider appeal of CMS open data, the CMS Communications Group organized an "Ask Me Anything" or AMA session on the social platform reddit [31]. Participants included representatives of CMS (including three of the authors of this document), CERN IT and the CERN Library team. The AMA session presented an opportunity not only to discuss our views on open science as well as open data and data preservation but also to answer questions that the public may have about the data release. It was also an occasion to informally highlight that we, the data providers, are aware that it isn't easy to use the data and, while we do not have many resources to help with projects developed on top of them, we still think releasing real LHC data is an exercise worth doing.

This engagement session proved to be very successful: in quantative terms, compared with the regular traffic, there was a five-fold increase in hits to the CERN Open Data Portal following the AMA [32].

5.3. Contacting the team

Although limited human resources prevent large-scale support for users, it was felt that a means for contacting the core team behind the CERN Open Data Portal (including CMS representatives) should be available, with support provided on a best-effort basis. Therefore, a mailing list (opendata-support@cern.ch) with core members was set up and advertised via the portal itself.

A variety of people have contacted the mailing list not only presenting ideas and feedback for enhancing the portal (particularly on the usability/accessibility front) but have also wishing to collaborate on the project.

6. Usage

The CERN Open Data Portal was visited by 82000 distinct users during the month after the launch. Of these users, around 600 downloaded data files over HTTP, 5000 read the "About" pages, 21000 viewed the collections, 16000 used the event display, 3000 used the histogram application, 21000 viewed the records, and 10000 used the built-in search. The dataset download was estimated in January 2015, and there were around 1000 access hits through XRootD from the VMs and around 200 direct downloads from the portal [32].

7. Impact and outook

The CMS data release has fostered collaboration and enabled common solutions in data preservation. It has been the *primus motor* for the CERN Open Data Portal and opened a way for development of further services in the domain, which are now being prototyped. The pioneering work required to define metadata for complex records such as primary and derived datasets and the heterogeneous items in the tools collection has been an excellent starting point for further development in the area. The release has also encouraged other HEP experiment to discuss and decide their own data release conditions.

The impact of the data release has been, in general, very positive. The CMS Collaboration is now evaluating the release process, and, if there is no unexpected negative impact to the collaboration, the data releases will become a regular procedure.

Acknowledgments
The authors and the CMS Collaboration are grateful for the support we have received from CERN-IT and the Invenio team, CERN-GS-SIS (the library team), CERN-PH-SFT (for the CMS-specific CernVM image), the DPHEP collaboration, as well as the US NSF and DOE and the QuarkNet collaboration.

References

[1] CMS Collaboration 2008 The CMS experiment at the CERN LHC *JINST* **3** S08004 http://cern.ch/cms
[2] CMS Collaboration 2012 Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC *Phys. Lett. B* **716** 30
[3] Lassila-Perini K *et al.* 2014 Implementing the data preservation and open access policy in CMS *J.Phys.Conf.Ser.* **513** 042029
[4] http://opendata.cern.ch
[5] Hinzmann A 2011 Tools for Physics Analysis in CMS *J.Phys.Conf.Ser.* **331** 032042
[6] http://cms-sw.github.io/index.html
[7] http://linux.web.cern.ch/linux/scientific5/
[8] CMS Collaboration 2014 Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD) *CERN Open Data Portal* DOI: 10.7483/OPENDATA.CMS.B8MR.C4A2
[9] CMS Collaboration 2014 Electron primary dataset in AOD format from RunB of 2010 (/Electron/Run2010B-Apr21ReReco-v1/AOD) *CERN Open Data Portal* DOI: 10.7483/OPENDATA.CMS.PDY4.7H2H
[10] Rodriguez Marrero A 2014 Muons and electrons in PAT candidate format derived from /Mu/Run-2010B-Apr21ReReco-v1/AOD primary dataset *CERN Open Data Portal* DOI: 10.7483/OPEN-DATA.CMS.RJW2.QP44
[11] Rodriguez Marrero, A. (2014). Muons and electrons in PAT candidate format derived from /Electron/Run-2010B-Apr21ReReco-v1/AOD primary dataset *CERN Open Data Portal* DOI: 10.7483/OPEN-DATA.CMS.HHTK.9FS2
[12] Cecire K *et al.* 2014 The CMS Masterclass and Particle Physics Outreach *EPJ Web Conf.* **71** 00027 http://www.physicsmasterclasses.org
[13] http://creativecommons.org/publicdomain/zero/1.0
[14] http://invenio-software.org, http://urn.fi/URN:NBN:fi-fe2014070432236
[15] McCauley T 2014 Software to extract data in csv format from a CMS primary dataset *CERN Open Data Portal* DOI: 10.7483/OPENDATA.CMS.RB4W.3ZK9
[16] Rodriguez Marrero A 2014 Two-lepton/four-lepton analysis example *CERN Open Data Portal* DOI: 10.7483/OPENDATA.CMS.QXY9.X47P
[17] http://cernvm.cern.ch
[18] http://xrootd.org
[19] http://opendata.cern.ch/visualise/events/CMS
[20] Hategan M, McCauley T and Nguyen P 2012 A browser-based event display for the CMS experiment at the LHC *J.Phys.Conf.Ser.* **396** 022022
[21] https://github.com/deanm/pre3d
[22] Alverson G *et al.* 2012 iSpy: a powerful and lightweight event display *J. Phys. Conf. Ser.* **396** 022002
[23] https://github.com/inveniosoftware/invenio-previewer-ispy
[24] https://github.com/cms-outreach/ispy-webgl
[25] http://opendata.cern.ch/visualise/histograms/CMS
[26] http://d3js.org/
[27] http://www.flotcharts.org/
[28] http://press.web.cern.ch/press-releases/2014/11/cern-makes-public-first-data-lhc-experiments
[29] http://cms.web.cern.ch/news/cms-releases-first-batch-high-level-lhc-open-data
[30] https://twitter.com/cern/status/535443404755566592
[31] http://redd.it/2nxwkb
[32] T. Simko, private communication