Available online at www.sciencedirect.com

ScienceDirect

Nuclear and Particle Physics Proceedings 273–275 (2016) 1229–1232

www.elsevier.com/locate/nppp

Preparations for the public release of high-level CMS data

Kati Lassila-Perini

Helsinki Institute of Physics, Helsinki, Finland

Alicia Calderon, Ana Y. Rodriguez-Marrero

Instituto de Fisica de Cantabria (UC-CSIC), Santander, Spain

David Colling, Adam Huffman

Imperial College, London, United Kingdom

Achintya Rao

University of the West of England, Bristol, United Kingdom

Thomas McCauley

University of Notre Dame, Notre Dame, Indiana, USA

(on behalf of the CMS Collaboration)

Abstract

The CMS Collaboration, in accordance with its commitment to open access and data preservation, is preparing for the public release of up to half of the reconstructed collision data collected in 2010. Efforts at present are focused on the usability of the data in education. The data will be accompanied by example applications tailored for different levels of access, including ready-to-use web-based applications for histogramming or visualising individual collision events and a virtual machine image of the CMS software environment that is compatible with these data. The virtual machine image will contain instructions for using the data with the online applications as well as examples of simple analyses. The novelty of this initiative is two-fold: in terms of open science, it lies in releasing the data in a format that is good for analysis; from an outreach perspective, it is to provide the possibility for people outside CMS to build educational applications using our public data. CMS will rely on services for data preservation and open access being prototyped at CERN with input from CMS and the other LHC experiments.

Keywords: Open data, data preservation, outreach, education, open science, CMS, LHC

1. Introduction

CMS (Compact Muon Solenoid) [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 over 300 published papers describing searches for SUSY and exotica, measurements of QCD, electroweak, top, forward, heavy-ion, and B physics, as well as discovery of the Higgs boson [2].

Email address: thomas.mccauley@cern.ch (Thomas McCauley)

http://dx.doi.org/10.1016/j.nuclphysbps.2015.09.195

In recognition of the importance of data preservation, re-use, and open access, CMS has approved a policy (found via [3]) that defines the experiment's approach. The policy covers many levels, from a commitment to publication in open-access journals, to release of reconstructed data, and the preservation and release of software and documentation needed for reconstruction and analysis.

CMS has prepared and released a small, selected amount of data for use in education and outreach. These datasets were reduced to the level of four-vectors and contain $J/\psi, \Upsilon, W$, and Z candidates as well as general two-muon and two-electron events [4]. 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). These data have formed the core of the successful masterclass program aimed at high-school students around the world [5].

Moreover, CMS has approved the release of a large set of reconstructed data for public use. This dataset is around 30 TB of 2010 proton-proton collision data at 7 TeV (tens of pb^{-1}) in CMS Analysis Object Data (AOD) format [6]. This contribution describes the preparation and release of this dataset in an effort to maximise its utility to the public.

2. CMS public data release

Previous datasets released for education and outreach were prepared for that use-case. The data files contained reduced, easily-accessible and easily-understood content. In several cases, the data were prepared for particular exercises. A measurement of the W^+ to W^- ratio and the invariant mass of spectrum of dimuons are two examples.

The next release will contain fourteen primary datasets. The content of each of these datasets is the result of the application of particular selection criteria. The dataset names, such as MinimumBias, Mu, Electron, and MultiJet (to give a few examples), indicate the result of the desired selections.

The CMS AOD format contains information needed for an analysis such as physics objects (muons, electrons, jets, etc.), tracks with associated hits, calorimetric clusters with associated hits, vertices, trigger information, and data needed for further selection and identification criteria for the physics objects.

However, one of the challenges the public will face when confronted with the large complex dataset contained in the next release, from one of the largest and most complex scientific instruments in the world, is that one needs specialised physics and computing knowledge in order to extract meaningful results. One needs a particular software environment to even read the data. The question is: how does the public get value from the open data? The partial answer is to initially focus on an already-proven use-case: education and outreach.

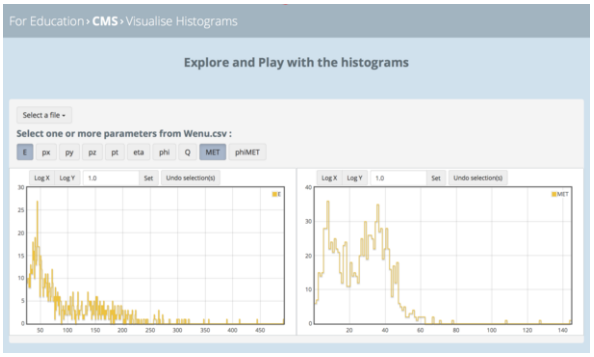

Figure 1: A screenshot of the prototype histogram application available via the open data portal.

3. Open data portal

CERN, in collaboration with CMS and the other LHC experiments, is preparing an open data portal [7] from which the data and tools for the public will be available. The portal is divided into two main areas, reflecting two different levels of access and complexity. The idea is to include and build upon the previous success of public data in education and outreach but also to include the possibility for more in-depth, complex analysis. The portal is therefore divided into two sections: "For Education" and "For Research".

3.1. Education

The education section includes "derived" datasets, which include data already prepared for masterclasses as well as datasets reduced to the level of four-vectors derived from the primary datasets. There is also access to a browser-based event display [8]. An interactive, online plotting tool is available as well: users can examine the information found in reduced datasets via histograms of distributions of kinematical parameters. An example can be seen in Figure 1.

3.2. Research

The "Research" section includes links to and documentation about the primary datasets. An example of a data record for the Mu dataset can be seen in Figure 2.

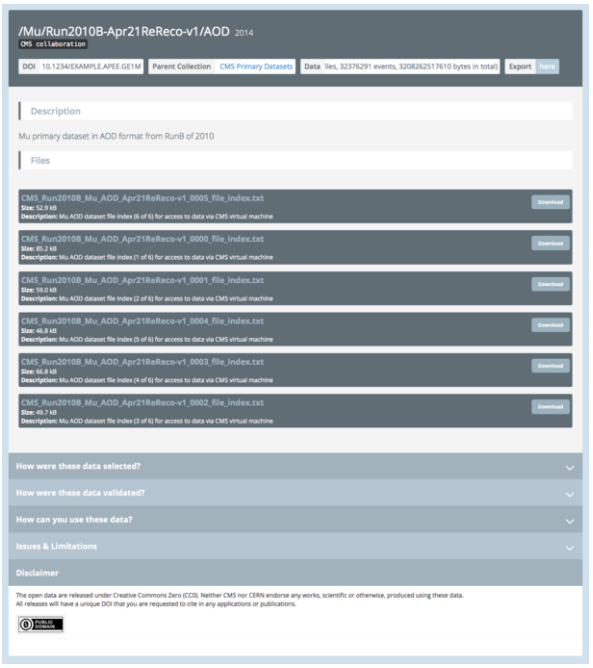

Figure 2: A screenshot of a data record for the ”Mu” primary dataset available in the open data portal.

A data record includes descriptions of the data selection and validation as well as information on instructions for analysis and limitations on usage.

However, in terms of usage, one needs a compatible computing environment in order to access the AOD in primary datasets in a meaningful way. To facilitate this access CMS has prepared a virtual machine (VM) with the appropriate software environment (in this case Scientific Linux CERN 5 [9]) with an appropriate release of CMS software (CMSSW [10]). Within the environment, access to example analysis code is made available and access to the data itself is available via XRootD [11].

The open-source examples include code for implementing a simple selection of dimuons from the Mu sample and generating the output of the four-vector information to a CSV file and code for conversion of events into the format suitable for viewing in the event display. Also included is code for doing a simple analysis of Z decays to two leptons and ZZ decays to four leptons; 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. An example two-lepton invariant mass spectrum from this analysis can be seen in Figure 3.

A representative sample of the each of the primary

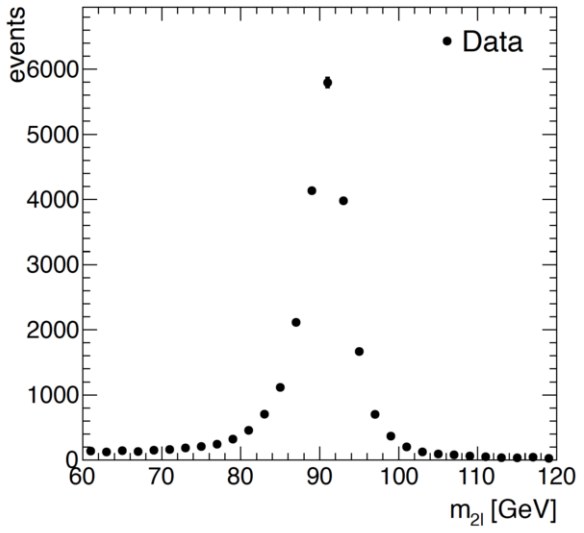

Figure 3: A example invariant mass spectrum that can be constructed from the open CMS data with the example analysis code.

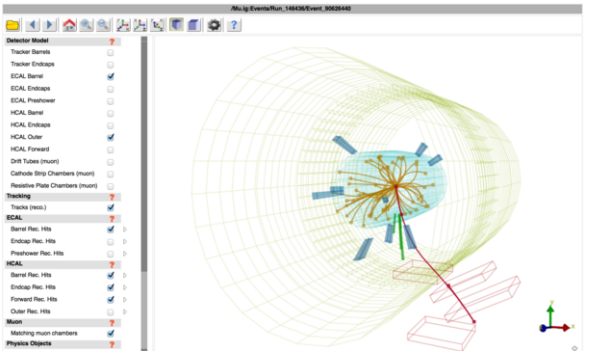

Figure 4: A screenshot of an event from the ”Mu” primary dataset in the browser-based event display available in the open data portal.

datasets is available for viewing in the browser-based event display. A screenshot of the display is shown in Figure 4.

3.3. *Invenio and CERN support*

The portal uses Invenio digital library software [12] as its base. A familiar example (to physicists) of an application that uses Invenio is the CERN Document Server (CDS) [13]. Invenio provides document organisation, search capability, and handling of metadata. CMS will rely on CERN support and services for legacy data storage, access to and distribution of the data, and security and bandwidth restrictions for public access.

3.4. Data re-use

The data will be released under the Creative Commons CC0 waiver [14], essentially releasing it into the public domain. The data will be identified with persistent digital object identifiers (DOIs) and it is expected that third parties will cite the CMS public data through these identifiers. Any possible publications based on public data by members of the CMS collaboration will be regulated.

4. Outlook

The CMS experiment, in recognition of its commitment to data preservation and open access as well as to education and outreach, has agreed to release up to half of its collected data from 2010. In addition, and in collaboration and with the support of CERN, it is preparing an open data portal from which the data will be available. CMS has also prepared several aids to the public in order to aid in usage and exploration. This includes online applications such as an event display and histograms of event information as well as a virtual machine with the CMS software environment with access to the data. Example analysis code as well as documentation is provided.

This is an excellent opportunity to understand what public data implies in terms of the external users, the impact on the experiments and the public, and resources needed. It is also an excellent opportunity to demonstrate to the public and to funding agencies that commitments concerning data preservation and open access are being taken seriously.

5. Acknowledgements

We wish to thank collaborators on CMS and gratefully acknowledge the help of CERN IT and Library Services. We also wish to thank the organisers of ICHEP. This work was partially supported by the US National Science Foundation.

References

[1] CMS Collaboration, The CMS experiment at the CERN LHC, JINST 3 (2008) S08004. http://cern.ch/cms
[2] CMS Collaboration, Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, Phys. Lett. B716 (2012) 30.
[3] K. Lassila-Perini *et al.*, Implementing the data preservation and open access policy in CMS, J.Phys.Conf.Ser. 513 (2014) 042029.
[4] http://cms.web.cern.ch/content/cms-public-data
[5] K. Cecire *et al.*, The CMS Masterclass and Particle Physics Outreach, EPJ Web Conf. 71 (2014) 00027 and T. McCauley, The QuarkNet CMS masterclass: bringing the LHC to students, these proceedings.
[6] A. Hinzmann, Tools for Physics Analysis in CMS, J.Phys.Conf.Ser. 331 (2011) 032042.
[7] http://opendata.cern.ch
[8] M. Hategan *et al.*, A browser-based event display for the CMS experiment at the LHC, J.Phys.Conf.Ser. 396 (2012) 022022, http://cern.ch/opendata-ispy
[9] http://linux.web.cern.ch/linux/scientific5/
[10] http://cms-sw.github.io/index.html
[11] http://xrootd.org
[12] http://invenio-software.org, http://urn.fi/URN:NBN:fi-fe2014070432236
[13] http://cdsweb.cern.ch
[14] http://creativecommons.org/publicdomain/zero/1.0