PROCEEDINGS OF SCIENCE

Open Data at CMS: Status and Plans

Thomas McCauley* **on behalf of the CMS Collaboration**
University of Notre Dame (US)
E-mail: thomas.mccauley@cern.ch

The CMS Experiment at the LHC has released many large datasets of proton-proton collision data as well as simulation to the public as part of its commitment to data preservation and open access. The collision data released totals around 2 fb^{-1} at 7 TeV and nearly 12 fb^{-1} at 8 TeV.
Data preservation describes the efforts to not only preserve the data itself but the conditions in which it can be analyzed. This requires archiving and documenting information such as run environment and conditions and tools such as analysis software and workflows. Open access describes the efforts to share these tools and information with the public, including educators and researchers.
These efforts present many challenges, partly due to the complexity and size of the CMS datasets and also to the knowledge required to make meaningful use of them. We describe here how we dealt with these challenges and the usage of CMS open data not only in public education and engagement but also in fundamental research. Furthermore we describe plans for future releases of data.

7th Annual Conference on Large Hadron Collider Physics - LHCP2019
20-25 May, 2019
Puebla, Mexico

*Speaker.

PoS(LHCP2019)260

1. Introduction

During Runs 1 and 2 of the Large Hadron Collider (LHC) over the years 2010 to 2018 the CMS Experiment [1] has collected over 6 fb^{-1} of proton-proton collision data at $\sqrt{s}$ = 7 TeV, over 23 fb^{-1} at $\sqrt{s}$ = 8 TeV, and 163 fb^{-1} at $\sqrt{s}$ = 13 TeV. With these data CMS has conducted a broad physics program leading perhaps most famously to the discovery of the Higgs boson [2].

The data collected by CMS are a unique scientific resource, impossible to produce again, and are therefore worthy of preservation. These data, and the results produced from their analysis, comprise a large part of the scientific legacy of CMS. In recognition of this legacy, and in the benefits of the data to public education and engagement, and to the larger research community, CMS has drafted, adopted, and carried out a data preservation and open access policy.

Data preservation describes the efforts to preserve the data as well as the conditions in which it can be analyzed and open access means making the data and conditions available and useable to the public. Both concepts are interdependent as a dataset can't be used (and reused) unless it and the conditions for its use are preserved. Data used are data preserved. A comprehensive perspective on open data and reproducibility may be found here [3].

This paper begins by describing the current fulfillment of the CMS data preservation and open access policy. It continues with a discussion of the challenges faced and how they were tackled, with specific examples. We end with a discussion of future plans.

2. Open data policy and current releases

Data preservation and open access policy in CMS [4] includes a commitment to publish 50% of collision data after three years and after ten years to release up to 100%. The data are released under an open license, the Creative Commons CC0 waiver [5], which essentially releases to the public domain. There have been several previous releases of CMS collision data since the first in 2014. The current releases of proton-proton collision data in terms of year, integrated luminosity, and center-of-mass energy are itemized below.

- 2010: 32 pb^{-1} at $\sqrt{s}$ = 7 TeV
- 2011: 2.3 fb^{-1} at $\sqrt{s}$ = 7 TeV and simulation
- 2012: 11.6 fb^{-1} at $\sqrt{s}$ = 8 TeV and simulation

CMS and the other LHC experiments have adopted similar concepts of levels of access to data. The lowest and simplest level is Level 1, which describes data directly related to publications such as original figures and values from plots and tables. Level 2 describes simplified data formats suitable for education and outreach. The next level of complexity is Level 3, which describes "analysis level" reconstructed data and simulation and associated software. It is at this level that the one would find the particle physicist conducting analysis. Finally, and at the highest level of complexity is Level 4, which describes raw, unreconstructed data and associated software. CMS open data efforts mostly focus on Levels 2 and 3 although an initial Level 4 release is described here.

All open datasets, software, and documentation are available via the CERN Open Data Portal [6], a collaborative effort between the CERN Information Technology and CERN Scientific Information Services teams and the experimental collaborations who release open data on it. The CERN Open Data Portal is built with the Invenio [7, 8] library management software framework. Products such as datasets, software, and documentation are shared under open licenses and are issued digital object identifiers (DOIs) so that they are citeable. Backend data storage is provided by EOS [9].

3. Challenges, solutions, and use-cases

There are several challenges that one must face in making such large and complex datasets public and useable. The data do not exist on their own but are just one part of an analysis "ecosystem". Therefore analysis software and workflows as well as metadata such as detector conditions and corrections need to be preserved and documented as well. Finally, a considerable amount of combined physics and software knowledge is needed in order to make meaningful use of the data.

These challenges are met in several ways. A general solution is to provide for different levels of knowledge and expertise, from students up to professional researchers, roughly corresponding to the levels of access described in the previous section. For students it is perhaps more appropriate to provide simplified datasets, ones that are in less complex, easier-to-use formats and that also contained simplified information. An example would be a text-based format which lists the four-vector information for physics objects such as muons from already-selected events. Such files are used in education programs such as masterclasses. At various levels of complexity example analysis code is available and is publicly accessible either on the CERN Open Data Portal or on a publicy accessible software repository like GitHub (in many cases on both). For more advanced users software environments such as virtual machines and containers are provided. In these environments one has access to datasets, experiment software framework, and conditions. Finally, it should almost go without saying that comprehensive, clear, and maintained documentation is essential.

A more detailed discussion of the current uses of CMS open data will illustrate further how the challenges are met. The first and most enduring use of open data has been in the International Masterclasses [10], particle physics analysis exercises with literally global reach aimed at the high-school level. The datasets used in these exercises are derived, simplified datasets such as those found in "Event files for CMS masterclass exercise 2014." [11]. In addition there is an active effort to create open source exercises in multiple languages using simplified formats analyzed using Jupyter notebooks [12], interactive analysis environments run in a browser. These exercises have been used in teacher training in Finland [13] and have been partly used and developed in the course of the CERN High School Teacher program. More materials related to both programs may be found on GitHub [14].

CMS has also provided many analysis examples, including one [15] based on the discovery of the Higgs boson using data collected in 2011 and 2012. This is a well-documented analysis available at different levels of complexity, from simply making the four-lepton invariant plot seen in Figure 1, up to replicating the original analysis.

More advanced users may wish to have full access to CMS open data in its native format along with CMS software and conditions. In this case CMS has provided virtual machine images

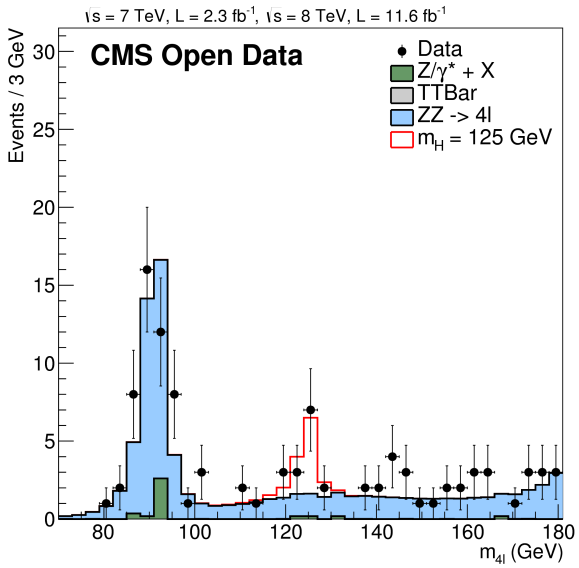

Figure 1: Invariant mass plot of events with four muons passing the selections found in the "Higgs-to-four-lepton analysis example using 2011-2012 data." [15]

containing the appropriate and necessary software and conditions. From here, the user has a complete environment in which to analyze CMS data at the level of the physicist. A lighter-weight alternative to a full virtual machine image is a Docker [16] container in which one has a self-contained environment that shares the kernel and resources of the host system on which it is run. Docker containers with the CMS analysis environment [17] are available from versioned public repositories [18]. Among the benefits of using a container is that their use can aid in reusability and reproducibility. One can fetch environment and analysis code and run in one command; an example can be found here [19].

Advanced users have made use of the data and tools for research purposes. This research includes searches for new particles [20, 21], jet and QCD studies [22, 23], and machine learning studies [24, 25, 26]. Furtherfore, and very usefully, some of the research users have provided valuable and detailed feedback and perspectives [27, 28] on the challenges one faces when embarking on research with CMS open data.

4. New dataset releases and future plans

The latest release of data reflects the ever-growing application of machine learning (ML) techniques to challenges in high-energy physics and the increasing collaboration of physicists and the data science and ML community [29]. The challenges for ML application in high-energy physics include particle identification, tracking and mitigation of pile-up.

One dataset [30] derived from Run 2 proton-proton simulation data at $\sqrt{s} = 13$ TeV focuses on the problem of b-jet tagging in which the b quarks come from the decay of a Higgs boson, a channel

which suffers from a formidable QCD background. Further datasets are devoted to the challenges of top quark identification [31] and to studying the flavor content of jets [32]. Another dataset [33] 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. All of the ML datasets are released with extensive documentation on their content, how to use them, and how to reproduce them with modified content. More information on the state of ML datasets and their use can be found in this contribution [34] to this conference.

As stated before, part of CMS' data preservation and open access policy is to release 100% of data within ten years of collection. As a fulfilment of this commitment, the remainder of proton-proton collision data at 7 TeV center-of-mass collected in 2010 are included in this new release. In addition to these data, a small sample of raw data from each of the years 2010-2012 are also released. An example of which can be found here [35]. These samples will facilitate testing of the processing chain from raw to reconstructed data in the legacy environment and open the possibility for future reconstruction algorithm studies and represent the first public release of Level 4 data.

The CASTOR calorimeter [36] operated in the very-forward (-6.6 < η < -5.2) region of CMS during LHC Runs 1 and 2. Reconstructed data and simulation collected in 2019 at $\sqrt{s}$ = 0.9 and 7 TeV are included [37, 38, 39] in the new release, representing the first release of data exploring this region of kinematic phase space.

In addition to data CMS has also prepared and released instructions and examples on how to generate simulated events and how to analyze data in a Docker container [40], within which one has access to the CMS software environment and can analyse data. Finally there have also been improvements in searching through the simulated data and to discover the provenance of datasets.

Future plans for open data on CMS include a continuation of release policy, continuous maintenance and improvement of documentation and example analyses (responding always to valuable user feedback), and to further exploit the educational potential of CMS data.

5. Acknowledgements

We would like to thank the CERN Information Technology and CERN Scientific Information Services teams for providing resources and expertise to build and maintain the CERN Open Data Portal. We would also like to thank the organizers of the LHCP for a productive and informative conference.

References

[1] CMS Collaboration, *JINST* **3** S08004 (2008).

[2] S. Chatrchyan *et al.* [CMS Collaboration], *Phys. Lett. B* **716** (2012) 30 [arXiv:1207.7235 [hep-ex]].

[3] X. Chen *et al.*, *Nature Phys.* **15** (2019) no.2, 113 [DOI:10.1038/s41567-018-0342-2].

[4] CMS Collaboration (2018), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.7347.JDWH].

[5] `https://creativecommons.org/publicdomain/zero/1.0/`

[6] `https://opendata.cern.ch`

[7] J. Kuncar, L. H. Nielsen, T. Simko, `http://urn.fi/URN:NBN:fi-fe2014070432236`

[8] http://invenio-software.org

[9] A. Peters *et al.*, http://eos.web.cern.ch

[10] https://physicsmasterclasses.org

[11] T. McCauley (2014), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.SSYF.EGXW].

[12] https://jupyter.org

[13] https://icps.helsinki.fi/myabstract/2019/full.php?sid=edu-10006

[14] https://github.com/cms-opendata-education

[15] N. Z. Jomhari, A. Geiser, A. A. Bin Anuar (2017), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.JKB8.RR42].

[16] https://docker.com

[17] C. Lange, clelange/cmssw-docker: Release for Zenodo (Version v1.0). Zenodo. http://doi.org/10.5281/zenodo.3374808

[18] https://hub.docker.com/u/cmsopendata

[19] https://github.com/reanahub/reana-demo-cms-h4l

[20] C. Cesarotti, Y. Soreq, M. J. Strassler, J. Thaler, W. Xue, *Phys. Rev. D* **100** (2019) no.1, 015021 [arXiv:1902.04222 [hep-ph]].

[21] C. G. Lester and M. Schott, [arXiv:1904.11195 [hep-ex]].

[22] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, J. Thaler, *Phys. Rev. D* **96** (2017) no.7, 074003 [arXiv:1704.05842 [hep-ph]].

[23] A. Larkoski, S. Marzani, J. Thaler, A. Tripathee, W. Xue, *Phys. Rev. Lett.* **119** (2017) no.13, 132003 [arXiv:1704.05066 [hep-ph]].

[24] C. F. Madrazo, I. H. Cacha, L. L. Iglesias, J. M. de Lucas, [arXiv:1708.07034 [cs.CV]].

[25] M. Andrews, M. Paulini, S. Gleyzer, B. Poczos, [arXiv:1807.11916 [hep-ex]].

[26] M. Andrews *et al.*, [arXiv:1902.08276 [hep-ex]].

[27] http://opendata.cern.ch/docs/cms-the-future-is-open-2017

[28] M. Strassler, J. Thaler, *Nature Physics* **15** 725 (2019). https://doi.org/10.1038/s41567-019-0628-z

[29] K. Albertsson *et al.*, *J. Phys. Conf. Ser.* **1085** (2018) no.2, 022008 [arXiv:1807.02876 [physics.comp-ph]].

[30] J. Duarte (2019), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.JGJX.MS7Q].

[31] A. Di Florio, F. Pantaleo, M. Pierini (2019), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.N1IN.TQHD].

[32] K. Kallonen (2019), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.RY2V.T797].

[33] E. Usai, M. Andrews, B. Burkle, S. Gleyzer, M. Narain (2019), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.CHC3.5KPG].

[34] J. Duarte on behalf of the CMS Collaboration, these proceedings.

[35] CMS Collaboration (2019), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.W24L.SGYC].

[36] P. Gunnellini [CMS Collaboration], [arXiv:1304.2943 [physics.ins-det]].

[37] CMS Collaboration (2019), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.QJ68.VK85].

[38] CMS Collaboration (2019), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.1R58.OMBD].

[39] CMS Collaboration (2019), CERN Open Data Portal [DOI:10.7483/OPENDATA.CMS.HK0O.ZF9Q].

[40] `http://opendata.cern.ch/docs/cms-guide-docker`