PROCEEDINGS OF SCIENCE

HEP data for everyone: CERN open data and the ATLAS and CMS experiments

Thomas McCauley for the ATLAS and CMS Collaborations*
University of Notre Dame (US)
E-mail: thomas.mccauley@cern.ch

A cornerstone of good scientific practice is to make results available to the public. This is especially true for experiments at the LHC at CERN where public investment in fundamental research is significant and long-standing. As part of their commitment to open access and public engagement the ATLAS and CMS collaborations have made several large datasets available to the public. There are many challenges posed in presenting complex and high-level data to the public in an accessible and meaningful way. We describe the solutions to these challenges, part of which is the creation and use of the CERN Open Data Portal and the content found therein. Furthermore we describe the impact and future plans of the ATLAS and CMS open access efforts including future releases of data and accompanying educational material.

38th International Conference on High Energy Physics
3-10 August 2016
Chicago, USA

*Speaker.

PoS(ICHEP2016)335

1. The ATLAS and CMS experiments

ATLAS [1] and CMS [2] are the two general-purpose experiments at the Large Hadron Collider (LHC) at CERN. Both share a broad physics program including studies of Standard Model physics, searches for SUSY and other physics beyond the Standard Model, and measurement of the properties of the Higgs Boson, the particle discovered by ATLAS and CMS in 2012. The discovery of the Higgs was the culmination of Run 1 at the LHC and as of this writing the LHC has finished delivering proton-proton collisions for Run 2 in 2016, having delivered 41 fb^{-1} of integrated luminosity at $\sqrt{s}$ = 13 TeV to each experiment.

2. Open Data

The data stored from the LHC collisions is on the order of hundreds of petabytes and the LHC experiments have committed to make some of this data available to the public. The motivation for doing so was expressed by the then-Director General of CERN, Rolf Heuer when the first public LHC data was released in 2014: "We hope these open data will support and inspire the global research community, including students and citizen scientists".

Reflecting their commitment to open-access ATLAS and CMS have both drafted and adopted open-access policies [3, 4]. Both policies use common notions of levels of access to the data where each increasing level reflects the increasing complexity and information provided and the difficulty in delivering the data in a meaningful and useful way. Level 1 describes data directly related to publications. For example, the numerical data used in the production of a plot. Level 2 describes simplified data suitable for education and outreach. Level 3 describes "analysis-level" reconstructed data, simulation, and software, *i.e.* the data and tools needed and used by the physicists for analysis. Finally, Level 4 describes the raw data. These policies also include details on embargo periods, licensing, and data reuse.

The first and most enduring use of public data from the LHC is education. One example are the masterclasses [5]: simplified analyses of LHC data aimed at the high-school level. The success of such programs as the masterclasses was one of the favorable factors considered by the LHC experiments when further data releases and open-access policies were discussed.

2.1 Challenges

In order to make it useful and useable for the public several challenges have to be overcome. Firstly, the datasets involved are large and complex. Datasets of terabytes and petabytes in size present problems of storage, distribution and processing. Furthermore the datasets reflect the complexity of the experiments from which they are produced. In order to read, process, and analyse the data large custom-made software frameworks are required, the use of which requires not-inconsiderable software experience. Finally there is the physics knowledge required to properly analyze the data. In an experiment those analyzing the data either have or are working towards a PhD in physics. This is clearly not the average member of the public.

The details of how ATLAS and CMS responded to these challenges can and do differ but in general the common approach is to provide data and tools at different levels of knowledge and

expertise required. This includes preparing and providing simplified datasets, detailed and comprehensive documentation, example analysis code, easy-to-use and easy-to-access applications for analysis and visualization via a web browser, and finally, for more advanced users, virtual machines with the necessary software environment and access to the more complex analysis-level datasets.

3. Data releases

3.1 CERN Open Data Portal

CERN, in collaboration with the LHC experiments, has prepared the CERN Open Data Portal (CODP) [6] from which the data and tools for the public are 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: "Education" and "Research". Datasets are distinguished as either "primary" or "derived" (roughly falling into Level 3 and Level 2 categories, respectively)

The CODP is built with Invenio [7] digital library software which provides document organisation, search capability, and handling of metadata. The CODP relies on CERN support and services for legacy data storage, access to and distribution of the data, and security and bandwidth restrictions for public access. All four LHC experiments use the CODP to various extents; CMS at this point uses it most extensively and the following description of the CMS data and tools release makes frequent reference to the CODP.

3.2 CMS

The CMS open data and tools release intends to reach all levels of the public and therefore there is a spectrum of levels of access and complexity. In addition CMS addresses issues of data cataloging, validation, and preservation.

CMS has released two large datasets to the public. The first, in November 2014 was half of the 2010 proton-proton collision data at $\sqrt{s} = 7$ TeV, 27 TB in size, equivalent to an integrated luminosity of tens of pb^{-1}. The second, in April 2016, was half of 2011 proton-proton collision data at $\sqrt{s} = 7$ TeV, 100 TB in size, equivalent to around 2.5 fb^{-1}, along with 200 TB of Monte Carlo samples. These datasets correspond to Level 3 described in Section 2 and are released in CMS AOD (Analysis Object Data) [8] ROOT [9] format. The data is released under the Creative Commons CC0 waiver [10], essentially releasing it into the public domain. In addition the data is identified with persistent digital object identifiers (DOIs) and it is expected that third parties will cite the CMS public data through these identifiers. An image of some of the information included in a primary dataset record is shown in Figure [1].

In order to analyze the primary datasets one needs CMS software, access to the data, and the correct software environment. These are all provided via a virtual machine particular for each data release. The datasets are stored in EOS [11] and can be downloaded directly or accessed remotely in the virtual machines via XRootD [12]. Extensive documentation includes trigger and detector condition information as well as analysis code examples. A set of benchmark analyses which can be repeated using the public data are also available. They reproduce chosen measurements from

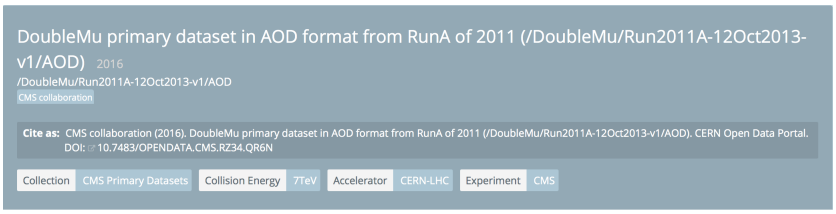

Figure 1: A screenshot of a dataset record showing the dataset name, citation, and DOI

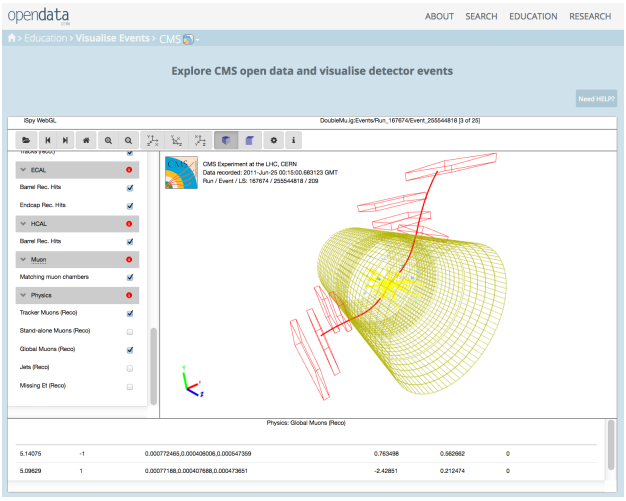

Figure 2: A screenshot of the browser-based event display for interactive visualisation of CMS events in the CERN Open Data Portal

published CMS papers and can therefore validate the public data and serve as examples for further analysis. Work is ongoing to provide a validation benchmark analysis for each primary dataset. This all comprises the content in the "Research" section of the CODP for CMS.

Content in the "Education" includes derived (*i.e.* reduced) datasets, corresponding to Level 2 in csv (comma-separated-value), JSON (JavaScript Object Notation), and CMS PAT (Physics Analysis Tool) [8] ROOT format. Also available in the "Education" section are two applications embedded in the CODP for easy interactive visualisation and examination of the datasets: an event display [13] and a histogram application [14]. A screenshot of the event display is shown in Figure [2].

3.3 ATLAS

Previous ATLAS data releases have included simplified datasets such as those used in the masterclasses and Monte-Carlo samples used for the ATLAS Higgs Machine Learning Challenge [15]. In July 2016 ATLAS released a large sample of data and tools with an initial focus on undergraduate and postgraduate students. Within this scope ATLAS provides access at three levels: from visualization, to analysis via a web browser, up to more complex analysis. At each level extensive documentation and examples are provided. The data, tools, and documentation are available via the ATLAS Open Data Page [16].

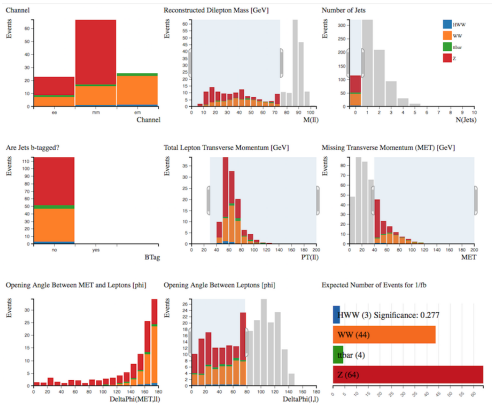

Figure 3: A screenshot of the interactive histogram tool found in the ATLAS Open Data Page.

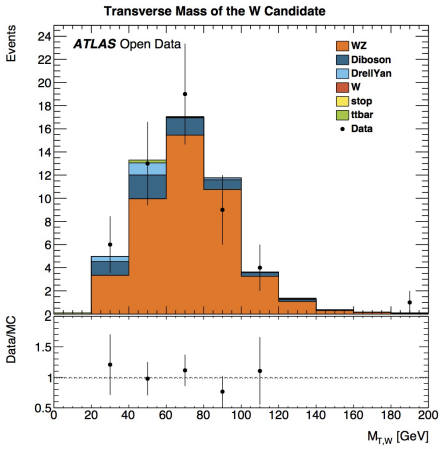

Figure 4: The distribution of the the W transverse mass which can be determined from an advanced ATLAS open data analysis.

The ATLAS dataset corresponds to 1 fb^{-1} of 2012 proton-proton collision data at $\sqrt{s} = 8$ TeV along with accompanying Monte Carlo. The data is in ROOT TTree format and contain electron/gamma and muon information. The first level of access with the ATLAS Open Data Page consists of an interactive histogram tool and a ROOT file browser. A screenshot of the histogram tool is shown in Figure [3]. At the next level interactive Jupyter [17] notebooks are available [18] where one can analyse the data interactively using the C++ or Python languages. Finally at the last, most-advanced, level [19] virtual machines with the environment and software necessary for analysis are available. An example analysis includes a WZ analysis where one looks for both a W boson candidate and Z boson candidate. An example result, the distribution of the transverse mass of the W, can be seen in Figure [4].

The intention is that the data and tools (including the virtual machine) at about 14 GB in size should all fit on a portable storage device such as a USB stick. This therefore allows for the data and tools to be distributed to the widest-possible audience where even perhaps online access is an issue. Data and virtual machines are also made available via CERN Open Data Portal.

4. Conclusions and future plans

Both ATLAS and CMS have released data to the public in a continuing effort to encourage public engagement, public education, and science. In order to maximize the potential of these data both experiments have prepared and released extensive tools and documentation and made them available online. Future plans include continuing release of data and improvement of data analysis tools and documentation. The goal will continue to be to ensure that as many as possible are able to use, learn from, and enjoy the public data.

5. Acknowledgements

We wish to acknowledge the support of the ATLAS and CMS Collaborations and their commitments to open access and gratefully acknowledge the support of CERN Scientific Information Services and the Invenio team. We also wish to thank the conference organisers for an especially engaging education and outreach program.

References

[1] ATLAS Collaboration, JINST 3 S08003 (2008), http://atlas.cern

[2] CMS Collaboration, JINST 3 S08004 (2008), http://cern.ch/cms

[3] ATLAS Collaboration (2014). ATLAS Data Access Policy. CERN Open Data Portal. DOI: 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ

[4] CMS Collaboration (2012). CMS data preservation, re-use and open access policy. CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.UDBF.JKR9

[5] http://physicsmasterclasses.org

[6] http://opendata.cern.ch

[7] http://invenio-software.org, http://urn.fi/URN:NBN:fi-fe2014070432236

[8] A Hinzmann, Tools for Physics Analysis in CMS, J.Phys.Conf.Ser. 331 (2011) 032042.

[9] R Brun, F Rademakers, ROOT - An Object Oriented Data Analysis Framework, Nucl. Inst. and Meth. in Phys. Res. A 389 (1997) 81, http://root.cern.ch

[10] http://creativecommons.org/publicdomain/zero/1.0

[11] http://information-technology.web.cern.ch/services/eos-service

[12] http://xrootd.org

[13] http://opendata.cern.ch/visualise/events/CMS

[14] http://opendata.cern.ch/visualise/histograms/CMS

[15] ATLAS Collaboration (2014). Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. CERN Open Data Portal. DOI: 10.7483/OPENDATA.ATLAS.ZBP2.M5T8

[16] http://atlasopendata.web.cern.ch

[17] http://jupyter.org/

[18] http://atlas-opendata.web.cern.ch/atlas-opendata/webanalysis/

[19] http://atlas-opendata.web.cern.ch/atlas-opendata/extendedanalysis/