PROCEEDINGS OF SCIENCE

ATLAS and CMS Data Release & Tools

Felix Socher* **on behalf of the ATLAS and CMS collaboration**
Institut für Kern- und Teilchenphysik, TU Dresden
E-mail: felix.socher@cern.ch

Science communication and outreach aim to engage the general public by making current research accessible. High energy physics (HEP) experiments collect large volumes of data. Analysis of these data is one of their key aspects. It is therefore desirable to make the data analysis process accessible to the general public. In doing so, the data and methods of HEP experiments become transparent to the outside world. ATLAS and CMS, two HEP experiments at the Large Hadron Collider, have implemented policies under which data and tools are to be made available to the public. The guidelines, tools, and data made available by the two collaborations will be discussed in this document.

Fourth Annual Large Hadron Collider Physics
13-18 June 2016
Lund, Sweden

*Speaker.

PoS(LHCP2016)112

1. Introduction

One of the key aspects of good scientific work is to communicate the findings of research to the general public. In high energy physics the general situation is that papers are produced for an informed audience while articles, physics briefings, videos and pictures address the general public. With research based on a large body of collected data it is tempting to use this data to produce more interactive devices for science communication. The motivation for doing so is manifold.

Enabling people to interact with actual data from the experiments is more engaging than traditional formats as the recipient is able to do something instead of just receiving information. Thus, the message is sent to the audience that it is invited to join in the scientific process and make its own discoveries instead of experiencing those of others. In addition, it makes the general working approach in a high energy physics experiment more transparent. This is an important point as it demystifies the way physicists work which may lead to a higher acceptance and regard for physics research. Last but not least, exercises that build upon actual data may provide motivating activities for students of physics thus preparing the next generation of particle physicists.

However, data should not be released as an uncommented package of raw information. It is important to consider the possible audiences and their preexisting knowledge. Such possible audiences are: the general public, high school students, university students, or students of physics in a graduate course. Depending on the intended audience the way of preparing the data, tools and accompanying documentation changes. For instance, university students are more likely to be able to program than the general public. Therefore, the data may be presented in a more complex way giving a richer experience while introducing requirements on prior knowledge. In contrast, material for a wide audience should not require prior knowledge but be accessible and enjoyable.

Both the CMS and ATLAS collaboration provide guidelines [1, 2] detailing the rules to be followed when making data available to the people outside the collaborations. Four levels are identified in these policies:

- Level 1 is published results, e.g. papers published in journals or conference notes. The actual data in these publications is made available in the form of tables and plots depicting results such as cross sections, selection efficiencies, and limits on model parameters. Analyses written in RIVET [3] fall into this category, too. The provided information may be used to test theoretical models or develop new ones. However, no direct interaction with the data on an event-wise basis is possible on this level.

- Level 2 signifies simplified data formats for outreach and educational use. These formats do not contain the full information available to members of the collaboration. They are rather geared towards demonstrating certain aspects of research in high energy physics in a simplified and easy to use environment. An example for level 2 data are event displays that allow for a visual inspection of individual collision events. Such data may be real or simulated collision data.

- Level 3 data is released in the actual data format used in the collaborations. These data formats are usually customised formats containing a high amount of detailed information on event properties. It is therefore necessary to supply a complete set of tools for data analysis.

Scientific use of these data is possible, but requires prior knowledge of high energy physics research and in most cases additional local computing resources. It can, however, be used to develop new exercises for more inclusive audiences.

- Level 4 signifies raw data as it is delivered by the detector. It is not regarded as sensible to release such data as it is very large in volume and needs a precise understanding of the machine at hand to extract sensible results. Therefore it is highly unlikely that such data will be made public apart from small samples for specific content-independent studies such as machine learning.

It is important to identify early on which level of data access is needed for a new exercise. For education purposes level 2 is the most interesting as it enables users with little prior knowledge to work in a simplified environment. However, level 3 data is also very important as it may be used for data preservation purposes enabling users outside the collaborations to redo analyses.

Access to the data has to be easy and efficient. To this end the CERN open data portal [4] has been developed. It started operations in November 2014 [5] and is a hub for the data released by the LHC collaborations. All exercises and data packages mentioned in this document are available there.

2. Available Tools and Data from CMS

The CMS collaboration offers a wide range of interactive material ranging from interactive event displays to the release of collision data in the actual format used by the collaboration. The list of exercises discussed here is not exhaustive but gives an overview of what is available. The individual exercises will be described in the following.

2.1 CMS HEP Tutorial

The CMS HEP Tutorial [6] was developed to introduce undergraduate students to fundamental concepts of data analysis in HEP experiments. The goal of this exercise is to implement a top pair cross section measurement using about 50 pb^{-1} of CMS collision data from the 2011 data run. The collision data is accompanied by simulated data describing top pair production, W boson production in association with jets, Drell-Yan processes and QCD. The data is stored in plain ROOT [7] tuples with minimal information on the measured objects, thus enabling simplified analyses. Instructions, the data and code for its analysis are supplied in a 30 Mb large tar ball. The exercise can be classified as level 2 and its target audience are undergraduate students with some prior knowledge of C++ and ROOT.

2.2 CMS Masterclasses

The CMS masterclasses [8] are a physics programme whose target audience is high school students from the age of 15 to 18. They are 1-day long exercises consisting of introductory HEP lectures and an interactive data analysis part. During the masterclasses students analyse event displays depicting CMS data. Analysis goals are determining the $W^+ : W^-$ ratio and creating a mass plot to search for the Z boson, Higgs boson, and other unknown particles. The dataset consists of 3000 events from CMS containing W, Z, J/Ψ, Υ, background and Higgs events.

The individual events are analysed using the iSpy-WebGl event display [9] and the CMS Instrument for Masterclass Analysis (CIMA). The iSpy-WebGl event display is able to display the geometry of the CMS detector as well as the individual objects in the event such as tracks and calorimeter entries in 3-dimensions. Information on the individual objects can be obtained by clicking on them. The program is compatible with any (modern) browser (e.g. Firefox 3.6+ or Internet Explorer 10+). Event data are either available online for download or offline via a DVD.

The CMS masterclasses support a string of languages: Chinese, Dutch, English, French, German, Hebrew, Hungarian, Italian, Japanese, Polish, Portuguese, Spanish, and Turkish.

2.3 CMS Level 3 Data

In addition to the level 2 contents discussed previously, the CMS collaboration has also published level 3 contents. Here, data preservation enters as a second aspect next to outreach and education. In 2014, the collaboration released half of its 2010 collision data (several tens of pb^{-1} requiring 30 TB in storage size). This release was followed by a second one in 2016 with a dataset of 2.5 fb^{-1} of 2011 collision data being released in conjunction with suitable simulated data. The storage size of this second release of collision data is about 100 TB and 200 TB for the simulated data. The license for the released data is CC0 [10], thus releasing the data into the public domain. It is provided in the AOD (Analysis Object Data) format which is the actual analysis format used by the CMS collaboration. Virtual machines containing the version 4.2.8 and 5.3.32 of the CMS software CMSSW may be used to analyse the data from 2010 and 2011, respectively. In addition, files specifying the detector conditions have been made available providing necessary metadata for data analysis. Using the tools available a simplified data format (PAT files) can be written which are analysable via python scripts using PyROOT [11].

The tools and data are not intended for novice users but rather for experienced particle physicists with considerable prior knowledge. Documentation on the usage of the tools provided is available but may be too sparse for someone new to the field to effectively use the tools and data. However, the available resources can be used to develop new content such as lab courses, e.g. the mentioned exercise using PAT files.

2.4 Future Plans

The CMS collaboration is planning to release half of the 2012 collision data plus matching simulated data later this year as a part of their ongoing programme to release half of the data of each year-long data taking period after a suitable embargo interval. The data and tools to be released will be at level 3, as for the 2010 and 2011 data. In addition, the development of simplified data formats enabling people with less prior knowledge to use the release data and tools is planned. Further efforts are ongoing to provide validation analyses for each primary dataset reproducing published results such as the reproduction of the di-muon invariant mass plot [12].

3. Available Tools and Data from ATLAS

The focus of the ATLAS collaboration has so far been on level 2 data in terms of interactive data-based exercises. Currently no release of level 3 data is planned.

3.1 ATLAS Kaggle Higgs Challenge

The ATLAS Kaggle Higgs Challenge [13] was a machine learning challenge that was held in 2014 on the Kaggle platform. The stated task was to efficiently classify events into di-tau decays of a Higgs boson or background using machine learning methods. The intended target audience was people with prior knowledge in data analysis but not necessarily with a physics background. The challenge created a high amount of attention and is one of the Kaggle competitions with the most participants to date.

Due to this success it was decided to host the dataset, tools, and instructions used in the Kaggle competition on the open data portal. Thus anyone is able to download the data and tools used and exercise their machine learning skills. Consequently, the available data have been used in workshops as exercises.

3.2 ATLAS Masterclasses

As in CMS the intended audience of the ATLAS masterclasses are high school students. The ATLAS masterclasses are a long standing physics education programme which has reached thousands of students over the last years. Two separate exercises exist, the ATLAS W path [14] and the ATLAS Z path [15]. Both exercises are stable, easy to use and have been improved continuously over the past years.

In the ATLAS W path, students determine the $W^+ : W^-$ ratio and hunt for the Higgs in $H \to WW \to l\nu l\nu$ decays. The dataset is comprised of 12000 events of collision data from the 2011 data taking. It is a mixture of preselected events with the individual parts being enriched in top pair, W, Z, QCD, and WW events. The dataset is geared towards educational use with the individual parts being mixed in ratios that are not observed in real collision data but are optimised towards being an enjoyable and satisfying exercise.

Individual events are analysed using MINERVA which is derived from the event display software ATLANTIS [16]. Students inspect events visually deciding whether they can be categorised as single W production, a $H \to WW$ candidate or background. By categorising a sufficient number of events the $W^+ : W^-$ ratio can be determined and a first glimpse of the Higgs boson in $H \to WW$ events can be seen.

In the ATLAS Z path, the goal is to select events with two and four leptons thus discovering Z, ZZ, and $H \to ZZ$ events. Using ATLAS data recorded in 2011 a set of 36000 events was produced containing the signal processes as well as background. As for the W path this dataset does not represent the situation encountered in nature but is tailored towards education. In addition to the Standard Model processes, a Z' contribution is added to the datasets which present an unexpected signal making the exercise more interesting. The visual inspection of the events is done using HYPATIA which is derived from the ATLANTIS event display.

Both programs require the user to install Java on machines used for exercises.

3.3 ATLAS Open Data dataset

A recent addition to the set of available exercises has been the ATLAS Open Data dataset [17]. It consists of collision data taken in 2012 with a sample size of 1 fb^{-1} together with matching simulated data representing Standard Model processes. The provided data has a combined size of

7 GB. It is provided in a simplified data format containing basic event information and kinematic, isolation and quality information of the leptons and jets found in the respective events. The data is analysed using python scripts relying on the PyROOT library. These python scripts provide the facilities for reading the files, example analyses as well as plotting facilities which make producing histograms straight forward and easy. A virtual machine containing the data and tools is provided having an overall size of about 14 GB. Due to its size and simplicity the ATLAS Open Data dataset is a suitable foundation for implementing new exercises such as lab courses or for use in lectures as visualising material.

The ATLAS Open Data dataset is supposed to be usable by anyone interested in trying out data analysis with ATLAS data. Therefore good documentation is important and a dedicated website accompanying the released data has been prepared to provide information and support for setting up the environment and using the available tools and data. The documentation will be expanded continuously to enable more and more people with little to no prior knowledge the possibility to profit from the ATLAS open data release. It may therefore also prove a valuable resource for physics students new to particle physics, e.g. in the course of a bachelor thesis.

3.4 Future Plans

Depending on the feedback and needs of the users further collision data may be released in the coming years. In the meantime the available analysis tools and documentation will be updated continuously to provide a more and more refined experience. The ATLAS collaboration has founded the ATLAS Outreach Data and Tools group to give continuity to such efforts. The group also supports and supervises the creation of new outreach based exercises.

4. Conclusions

Experiences made with the masterclasses have shown that interactive exercises motivate students very well to engage with particle physics concepts. They show in a clear and transparent way how data analysis and research as a whole in high energy physics work. It is therefore desirable to provide interactive exercises using real data collected by the collaborations to enhance the body of possible outreach resources.

Both the ATLAS and CMS collaborations have implemented guidelines regarding the release of collision data and tools. Many such exercises are already available and more will be made available in the future for the benefit of the general public. All of the exercises discussed in this document are available on the CERN open data portal.

The data releases by the LHC experiments have been a great success and have encouraged external users to use these data for their own purposes. They have also enabled the release and development of the CERN Open Data Portal as the access point to a growing range of data produced through the research performed at CERN.

References

[1] ATLAS Collaboration, *ATLAS Data Access Policy*, *CERN Open Data Portal* (2014) .

[2] CMS Collaboration, *CMS Data Preservation, Re-Use and Open Access Policy*, *CERN Open Data Portal* (2012) .

[3] A. Buckley, J. Butterworth, L. Lonnblad, D. Grellscheid, H. Hoeth, J. Monk et al., *Rivet User Manual*, *Comput. Phys. Commun.* **184** (2013) 2803–2819, [1003.0694].

[4] "Open Data Portal." http://opendata.cern.ch/.

[5] "CERN Makes Public First Data of LHC Experiments." https://home.cern/about/updates/2014/11/cern-makes-public-first-data-lhc-experiments.

[6] "CMS HEP Tutorial." http://opendata.cern.ch/record/50?ln=en.

[7] R. Brun and F. Rademakers, *ROOT - An Object Oriented Data Analysis Framework*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **389** (1997) 81 – 86.

[8] "CMS Masterclasses." https://quarknet.i2u2.org/page/cms-masterclass-2016-documentation.

[9] "ISpy-WebGL." http://ispy-webgl.web.cern.ch/ispy-webgl/.

[10] "CC0 1.0 Universal." https://creativecommons.org/publicdomain/zero/1.0/.

[11] "PAT-file based Z/ZZ Analysis with CMS 2010 Open Data." http://opendata.cern.ch/record/101. 10.7483/OPENDATA.CMS.QXY9.X47P.

[12] A. Geiser, I. Dutta, H. Hirvonsalo and B. Sheeran, *Validation Code For 2010 Mu and MuMonitor Datasets, Based on Di-Muon Mass Spectrum*, *CERN Open Data Portal* (2016) .

[13] "ATLAS Higgs Machine Learning Challenge." https://higgsml.lal.in2p3.fr/.

[14] "ATLAS Masterclasses W Path." http://atlas.physicsmasterclasses.org/en/wpath.htm.

[15] "ATLAS Masterclasses Z Path." http://atlas.physicsmasterclasses.org/en/zpath.htm.

[16] "Atlantis Event Display." http://atlantis.web.cern.ch/atlantis/.

[17] "ATLAS Open Data." http://atlas-opendata.web.cern.ch/atlas-opendata/.

PoS(LHCP2016)112