PROCEEDINGS OF SCIENCE

LHC experiments and their Open Data

Edgar Carrera Jarrin[a,*]
On behalf of the ALICE, ATLAS, CMS and LHCb collaborations
[a] *Universidad San Francisco de Quito,*
Diego de Robles S/N, Quito, Ecuador
E-mail: ecarrera@usfq.edu.ec

The CERN laboratory has released a new open data policy for the LHC experiments. Although this policy focuses on the common strategy for the release of research-quality data, it strengthens all aspects of open data, and reaffirms the commitment of all the LHC experiments towards open science. In this presentation we give a summary on the status of the open data efforts from each of the large LHC experiments as well as their plans and strategies under the new policy.

The Ninth Annual Conference on Large Hadron Collider Physics - LHCP2021
7-12 June 2021
Online

*Speaker

PoS(LHCP2021)334

1. Introduction

Recently, all the large experimental collaborations at the CERN's Large Hadron Collider (LHC) [1], namely ALICE, ATLAS, CMS and LHCb, have announced a new policy [2] in support of open science. The document makes a solid commitment to publicly releasing research-quality, scientific data as well as data for education and outreach. This policy and its implementation plan were endorsed by the Collaboration Boards of all large LHC experiments.

There are four main abstraction levels for the data produced by these experiments. The first one, called *Level-1*, relates to published results (e.g., articles in scientific journals) but also to numerical information, like likelihood functions used in a given study. *Level-2* is about datasets in simplified formats, commonly used for outreach and education. The heart of the new policy deals with *Level-3* data, which can be used for scientific research and are usually subject to latency/embargo periods and restrictions. The "raw" data, i.e., the unprocessed information coming out from the different experiments, are at the *Level-4* and are not considered useful for an external user.

Essentially, all LHC experiments share a common strategy for the release of Level-1, -2 and -4 data. Level-3 strategies, however, differ within a common ground. This common baseline is founded on the commitment of all large collaborations to release research-quality, calibrated data accompanied by provenance information, simulated samples, workflow examples and documentation. Experiments will release Level-3 data periodically after a prudent embargo time, which allows for the full understanding, processing and scientific exploitation of them. The aim is to start the public release of data within five years of the conclusion of the data taking period. Timelines and amounts for data releases are defined and approved by the management of each experiment and may be subject to special restrictions. However, full research-quality datasets will be made public at the end of each experimental collaboration.

All datasets prepared by the LHC experiments for public use are released using a common web site, the *CERN Open Data Portal* (CODP) [3]. There, datasets are shared under open licenses and are citable, i.e., they acquire a DOI [1] number. Software and some documentation is also provided under the CODP platform.

In these proceedings, we review the current status of the open data efforts in all the large LHC collaborations and their plans to put the new CERN open data policy in motion.

2. ALICE open data

Currently, 5% of Pb-Pb and 7% of 2010 proton-proton (p-p) collisions datasets in ESD [2] format from the ALICE [5, 6] experiment have been released on the CODP together with the needed software to analyze them via virtual machine and container technology. In addition, simplified datasets have been released in the TTree ROOT data format for outreach and education; they are mainly used for the ALICE Masterclass exercises, which are also available as web-based applications [7].

[1] Digital Object Identifier

[2] Event Summary Data (ESD): a format with detailed reconstruction output; Analysis Object Data (AOD): distilled subset from ESD [4].

In order to best align with the new CERN open data policy, the ALICE experiment expects to make public data releases periodically, after a latency period. The details of this timeline aligns with the new policy and general points outlined in Section 1.

As far as data products, at the time of this presentation, there is a campaign taking place in order to convert data collected during Run 1 and Run 2 [3], in the ESD and AOD [4] data formats, into a new AOD format, which is based on the ALICE's *O2 Project* [8]. The conversion is expected to be completed by the end of 2021. Therefore, it is quite likely that, starting in 2022, ALICE's datasets, for real collisions and simulations, will begin appearing in the CODP in this new AOD paradigm. Derived datasets with simpler structures (e.g., nanoAOD format) are also being considered to reduce the impact of the larger statistics expected in Run 3.

3. ATLAS open data

ATLAS [9] has released public data mainly for outreach and education [10]. Its traditional contribution to the Masterclass program is a good example of its commitment to providing open-access resources for education. Tools like histogram analyzers and Jupyter notebooks to explore real physics collisions have been used to train students at different levels.

ATLAS leads the LHC efforts on publishing full likelihood functions (Level-1 data). These functions encode all the necessary information to reinterpret [11] experimental results from a different theoretical perspective. These data are openly available on the open-access site *HEPData* [12].

Under ATLAS' renewed open data policy [13], which is in full agreement with the new CERN open data one, the Collaboration will prepare periodic releases for public use after an adequate embargo period. Timelines and data amounts will be defined in accordance with this policy. For Level-3 data, in particular, real collisions and simulated datasets will be released after casting them into the new PHYSLITE [14] format, which is a simplified scheme that contains calibrated objects and also information to compute systematic uncertainties. In addition, special datasets may be approved for release.

For research-oriented applications, the experiment is also offering the possibility to establish association programs between members of the Collaboration and external authors. Currently, these programs are the only way to get access to research-quality data and are considered case-by-case.

Lastly, the software framework needed to analyze these data will be provided, most likely, using container technology.

4. CMS open data

CMS [15] is at the forefront of the LHC open data effort for scientific research. The first batch of Level-3 data was released in 2014. To this day, most of the p-p collisions data from Run 1 have been released, including datasets that were used to discover the Higgs boson [16]. The latest release was the first batch of Run 1 Heavy Ions data at the end of 2020 [17]. A few other datasets have also been released for special studies. Additional Level-1 (numerical) information is also made available

[3] LHC's Run 1 refers to the data taking period between 2010 and 2012, while Run 2 refers to the period between 2015 and 2018. Run 3 will start in 2022.

[4] See footnote 2.

through open access platforms like *InspireHEP* [18] or *HEPData* [12]. More than 30 scientific articles [5] have been produced using CMS open data so far, with about half of them published in indexed journals.

CMS has also contributed to outreach and education by publishing datasets that have been used for the traditional Masterclass exercises and the preparation of material for teaching high school and college students.

The updated CMS open data policy [19] reaffirms the commitment to regularly publishing Level-3 datasets. CMS will usually make 50% of its data available 6 years after they have been taken and raise that availability to 100% within 10 years. Certain restrictions may apply and Collaboration Board approval is always needed, which can modify dates and sizes of the releases.

CMS will continue to release real collisions and simulated datasets in AOD format (for Run 1) and miniAOD and nanoAOD [20] (for Run 2 and beyond). First Run 2 p-p collisions datasets are expected to be released by the end of 2021 or early 2022. Currently, the software needed to decode this data is provided through virtual machines and containers. Database snapshot of detector conditions, needed to fully utilize the data, are provided through the CernVM File System (CVMFS) [21] or directly into the containerized applications. Data quality and luminosity information tools are also provided. Finally, an open data guide [22] is being developed along with the institutionalization of regular training events for external users.

5. LHCb open data

Aside from its contribution to the international Masterclass exercises, the LHCb [23, 24] collaboration released a small sample of simulated events in 2020, which were used externally to produce a scientific publication on quantum-inspired machine learning techniques [25].

In its revised policy [26], LHCb adheres to the new CERN LHC policy and the general timelines mentioned in Section 1. The experiment supports the principles of open access but expresses concerns about the very limited person-power currently available to devote to open data efforts.

LHCb expects to release real collisions data and simulations in DST and microDST formats. Associated software will be released using containers. Calibration tools will also become available publicly, with their usage at the users own responsibility.

6. Summary

The LHC experiments at CERN have announced a new, revitalized policy on open data. The new policy puts emphasis on the release of research-quality scientific data. LHC xperiments will start releasing Level-3 data within 5 years of the end of a given data taking period and will release full datasets at the end of the lifetime of each collaboration. All the experiments have made important contributions to outreach and education by publishing collisions data, simulations and tools needed for these activities. While strategies differ, and some restrictions may be imposed, all the large LHC experiments are committed to open data for research and outreach and education.

[5] This InspireHEP link currently gives a very loose pointer to some of these articles. It is not exact but could be used as a starting reference.

References

[1] L. Evans and P. Bryant, eds., *LHC Machine*, *JINST* **3** (2008) S08001.

[2] "CERN Open Data Policy for the LHC Experiments." https://cds.cern.ch/record/2745133, Nov, 2020.

[3] "CERN Open Data Portal." https://opendata.cern.ch/.

[4] G. Eulisse, A. Alkin, J.F. Grosse-Oetringhaus, P. Hristov, G.M. Innocenti and M.J. Kabus, *Data Analysis using ALICE Run 3 Framework*, *EPJ Web Conf.* **245** (2020) 06032.

[5] ALICE collaboration, *The ALICE experiment at the CERN LHC*, *JINST* **3** (2008) S08002.

[6] ALICE collaboration, *Performance of the ALICE Experiment at the CERN LHC*, *Int. J. Mod. Phys. A* **29** (2014) 1430044 [1402.4476].

[7] ALICE collaboration, "ALICE Masterclass." https://alice-web-masterclass.web.cern.ch/.

[8] "ALICE O2 Project." https://alice-o2-project.web.cern.ch/.

[9] ATLAS collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.

[10] ATLAS collaboration, "ATLAS Open Data." https://atlas.cern/resources/opendata.

[11] LHC Reinterpretation Forum collaboration, *Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2*, *SciPost Phys.* **9** (2020) 022 [2003.07868].

[12] "HEPData." https://www.hepdata.net/.

[13] ATLAS collaboration, "ATLAS Data Access Policy." http://opendata.cern.ch/record/413. 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ.

[14] ATLAS Software and Computing collaboration, *Columnar data analysis with ATLAS analysis formats*, *EPJ Web Conf.* **251** (2021) 03001.

[15] CMS collaboration, *The CMS Experiment at the CERN LHC*, *JINST* **3** (2008) S08004.

[16] CMS collaboration, "Observing the Higgs with over one petabyte of new CMS Open Data." https://opendata.cern.ch/docs/observing-higgs-over-one-petabyte-new-cms-open-data.

[17] CMS collaboration, "CMS releases heavy-ion data from 2010 and 2011." http://opendata.cern.ch/docs/cms-releases-heavy-ion-data.

[18] "InspireHEP." https://inspirehep.net/.

[19] CMS collaboration, "CMS data preservation, re-use and open access policy." http://opendata.cern.ch/record/415. 10.7483/OPENDATA.CMS.1BNU.8V1W.

[20] CMS collaboration, *A further reduction in CMS event data for analysis: the NANOAOD format*, *EPJ Web Conf.* **214** (2019) 06021.

[21] "CernVM-FS Documentation." https://cvmfs.readthedocs.io/en/stable/.

[22] "CMS Open Data Guide." https://cms-opendata-guide.web.cern.ch/.

[23] LHCb collaboration, *The LHCb Detector at the LHC*, *JINST* **3** (2008) S08005.

[24] LHCb collaboration, *LHCb Detector Performance*, *Int. J. Mod. Phys. A* **30** (2015) 1530022 [1412.6352].

[25] T. Felser, M. Trenti, L. Sestini, A. Gianelle, D. Zuliani, D. Lucchesi et al., *Quantum-inspired machine learning on high-energy physics data*, *npj Quantum Inf.* **7** (2021) 111 [2004.13747].

[26] LHCb collaboration, "LHCb External Data Access Policy." http://opendata.cern.ch/record/410. 10.7483/OPENDATA.LHCb.HKJW.TWSZ.

PoS(LHCP2021)334