PROCEEDINGS OF SCIENCE

Engaging Universities and Beyond - Open Data for Education and Research

Leonardo Toffolin[a,b] **for the ATLAS collaboration**

[a] *Department of Physics, University of Trieste,*
Via Valerio 2, Trieste, Italy
[b] *INFN Trieste - Gruppo Collegato di Udine,*
Dipartimento Politecnico di Ingegneria ed Architettura, University of Udine, Udine, Italy

E-mail: leonardo.toffolin@cern.ch

The ATLAS Open Data project has for many years successfully delivered open-access data, simulations, documentation and related resources for education and outreach use in High Energy Physics and related computer sciences based on data collected in proton–proton collisions at 8 TeV and 13 TeV at the LHC at CERN. These resources have found substantial application worldwide in universities, in schools, and in other public settings. Building on this success and in support of CERN's Open Data Policy the ATLAS experiment plans to continue to release 13 TeV data for educational purposes and – for the first time – also for research purposes. This contribution summarises the landscape of the existing ATLAS Open Data project: what resources are available and how have they been used, and what is planned for the future.

11th Large Hadron Collider Physics Conference 2023,
22-26 May 2023
Metropol Palace, Belgrade, Serbia

PoS(LHCP2023)297

1. Introduction

In 2020, CERN approved a policy document that empowers LHC collaborations to publish experimental data for open public access [1, 2]. Already previously, the ATLAS Collaboration [3] at the LHC [4] had set out in 2015 its Data Access Policy [5] which indicates the guidelines for a project regarding open access to ATLAS data by non-ATLAS members, with a special focus on education, training and outreach.

The aim of the ATLAS Open Data project is to provide data and analysis tools to high schools, undergraduate and graduate students from various institutions and universities, and to the general public with the aim of helping users understand how a high-energy physics analysis with real LHC data is performed. Also, the ATLAS Open Data project aims to meet high expectations from national governments and scientific funders for the effort made by the ATLAS experiment to make real data and educational resources available to the public.

Open Data represent a crucial cornerstone of science and their benefits to society are numerous. They enable public understanding and inform policy, and also promote citizen science and potentially increase innovation. They provide economic benefits and easier access to research, even advancement of research, and increase trust in research and scientists. They expose students and the general public to real ATLAS proton-proton collision data, as well as to methods, techniques and computational skills needed for data analysis, potentially raising their interest in particle physics and STEM (**S**cience, **T**echnology, **E**ngineering and **M**athematics) subjects. They support the long-term improvement of the scientific literacy of the public. They also provide resources to teach transferable skills in programming and analysis techniques.

2. ATLAS Open datasets

The ATLAS Open Data project comprises software tools and a subset of proton–proton collision data from the Large Hadron Collider (LHC) collected by the ATLAS detector and reconstructed into final physics objects such as leptons, photons and hadronic jets. Four levels of data-set uses are defined in total, including data and metadata from published results, dedicated subsets for outreach and education purposes, and reconstructed data for independent scientific analyses.

The ATLAS data available within the whole Open Data project belong to two different releases. The first release of the project correspond to approximately 1 fb^{-1} of LHC proton–proton data recorded in 2012 at a centre-of-mass energy of 8 TeV, called the ATLAS Open Data 2016 dataset [6]. It was launched in 2016 together with an analysis framework and seven different examples in Python. The dataset corresponds to approximately 15 million events. This is part of the data that allowed ATLAS to discover the Higgs boson [7]. For this reason, this fraction of the 2012 data has an important scientific, educational and historic value. Simulated data are also made available for various Standard Model and new physics signal models, for comparison with LHC data.

The second and most recent Open Data release comprises approximately 10 fb^{-1} of LHC proton–proton data recorded in 2016 at a centre-of-mass energy of 13 TeV, called the ATLAS Open Data 2020 dataset [8]. It corresponds to approximately 61 runs from the first four periods of the 2016 proton-proton data-taking, with a total amount of about 270 million proton–proton collisions. Similarly to the previous release, it was launched in 2020 together with an analysis framework.

Datasets are stored in a "flat ROOT ntuple format"; they include calibrated and simplified information about the reconstructed high-level objects and systematic uncertainties [6, 8].

Events are selected by applying several event-quality and trigger criteria, and classified according to the type and multiplicity of reconstructed objects; in particular, dataset identifiers, scale factor information, trigger (boolean) and kinematic variables are included. Elementary information about jets, leptons and missing tranverse energy is also stored in the ROOT ntuples for analysis purposes.

3. Software tools and infrastructure

In the ATLAS Open Data effort, different methods are provided to support online and offline clients with different level of analysis complexity, both in online Jupyter Notebooks [9] and downloadable virtual machines [10] preloaded with all the software needed for the analysis.

The 13 TeV ATLAS Open Data release is accompanied by a set of Jupyter notebooks that allow data analysis to be performed directly in a web browser, by integrating the ROOT framework with the Jupyter notebook technology (Fig. 1); this combination is called `ROOTbook`. As an example, with the help of the `ROOTbook` technology the diphoton invariant mass spectrum can be easily extracted from data, in order to give users the possibility to look at one of the Higgs boson golden channels, as reported in Fig. 1.

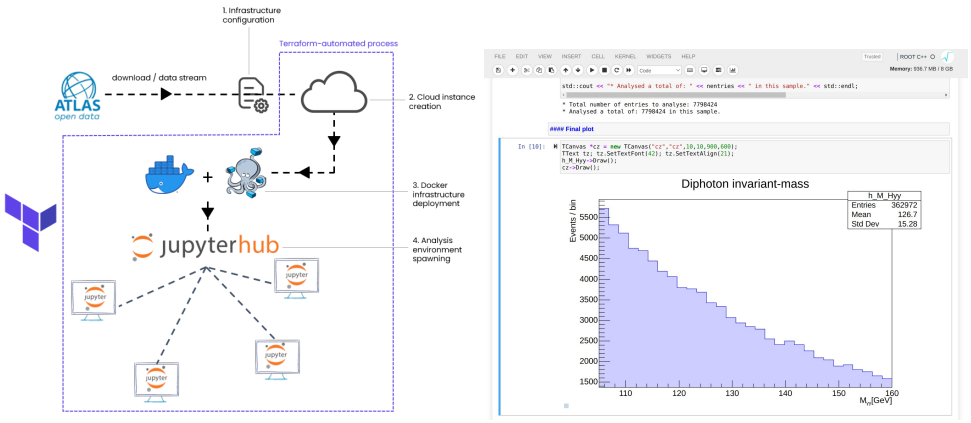

Figure 1: General structure of the Jupyter-based infrastructure (on the left); plot of the diphoton invariant mass from the 13 TeV Open Data release dataset (on the right).

Several notebooks with analysis examples and an interface to launch the framework mentioned above, are available using SWAN (**S**ervice for **W**eb-based **AN**alysis) and Binder executable platforms.

The ATLAS Open Data team is strongly active in developing an infrastructure to deploy reproducible educational data-analysis platforms at small or medium projects and institutions. This means downloading or streaming ATLAS Open Data resources to an external cloud, and then into institutional resources. Both the Jupyter notebooks and the local virtual machines approaches have pros and cons and could be taken into consideration depending on circumstances. The usage of Jupyter notebooks, without downloading anything, can be a good option if a solid internet

connection is available; local virtual machines with locally downloaded data and analysis code can be used in a completely offline environment as well; Docker containers [11] can combine the two approaches, presenting a notebook interface on top of a locally running web service.

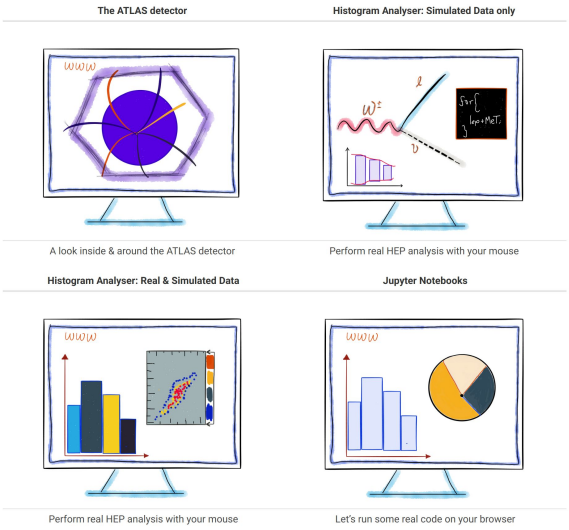

Figure 2: Simple scheme of the apps designed for the ATLAS Open data activities [12].

Moreover, the ATLAS Open Data team aims to deliver some web and desktop applications so that users can explore High Energy Physics (HEP) real and simulated data on their browser or computer (Fig. 2). For users, it is possible to explore a simulation of the ATLAS detector. Going more in-depth on the code, users can rely on several Jupyter notebooks examples to teach, run, modify and explore real analysis code without any complicated setup or installation.

4. Summary and future plans

Both the 8 TeV and 13 TeV ATLAS Open Data releases have been proved very successful for outreach activities. Many institutions used them over the years to engage high-school and university students. The Open Data campaigns provided great help in involving students from countries in various geographic locations, including some countries that are not part of the ATLAS Collaboration, throughout South America, the Middle East, Asia and Africa.

Still, there is a relevant number of key challenges to consider for future ATLAS Open Data activities for research and education. An upcoming data release is ramping up for 2024 to help students over the world to perform more analyses and learn more skills with the ever-evolving tool kit. The new Open Data campaign plans to include most of the LHC-Run 2 data.

In conclusion, data formats other than ROOT ntuples are begin considered. The so-called DOAD-PHYSLITE format, i.e. the standard data format which will be used for most ATLAS data analyses for LHC-Run 3 and HL-LHC, will be investigated as the starting point for the next ATLAS Open Data release. Great effort will be put to embrace new technologies with the software (Jupyter notebooks, Docker containers), which are not even used by all ATLAS analyzers at the state of art.

References

[1] Edgar Carrera Jarrin on behalf of the ALICE, ATLAS, CMS and LHCb collaborations, *LHC experiments and their Open Data*, PoS (LHCP2021) 334, https://cds.cern.ch/record/2802456.

[2] Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) Study Group, *CERN Open Data Policy for the LHC Experiments*, CERN-OPEN-2020-013, https://cds.cern.ch/record/2745133.

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

[4] Large Hadron Collider: https://home.cern/science/accelerators/large-hadron-collider.

[5] ATLAS Collaboration, *ATLAS Data Access Policy*, ATL-CB-PUB-2015-001, 2015, http://cds.cern.ch/record/2002139 (cit. on p. 3).

[6] ATLAS Collaboration, *Review of ATLAS Open Data 8 TeV datasets, tools and activities*, https://cds.cern.ch/record/2624572/.

[7] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716**, 1 (2012), doi:10.1016/j.physletb.2012.08.020 [arXiv:1207.7214 [hep-ex]].

[8] ATLAS Collaboration, *Review of the 13 TeV ATLAS Open Data release*, https://cds.cern.ch/record/2707171.

[9] ATLAS Open Data 13 TeV documentation, 13 TeV Open Data Jupyter Notebooks, http://opendata.atlas.cern/release/2020/documentation/notebooks/intro.html.

[10] ATLAS Open Data 13 TeV documentation, Virtual machines, http://opendata.atlas.cern/release/2020/documentation/vm/index.html.

[11] ATLAS Open Data, Docker containers, https://hub.docker.com/r/atlasopendata/root_notebook.

[12] ATLAS Open Data apps, http://opendata.atlas.cern/apps/.

PoS(LHCP2023)297