IL NUOVO CIMENTO **48 C** (2025) 124
DOI 10.1393/ncc/i2025-25124-0

Colloquia: IFAE 2024

ATLAS Open Data to engage the public in Education and Research(*)

Luca Clissa(1)(2) on behalf of the ATLAS Collaboration

(1) *Dipartimento di Fisica e Astronomia "Augusto Righi", Università di Bologna Bologna, Italy*
(2) *INFN, Sezione di Bologna - Bologna, Italy*

received 2 December 2024

Summary. — ATLAS Open Data is an initiative aimed at making the data, simulations, and documentary resources of the experiment accessible to a wide audience, in accordance with the CERN Open Data policy. The project has seen the release of numerous datasets of proton-proton collisions at center-of-mass energies of 8 TeV and 13 TeV collected at the LHC during Run-1 and Run-2, allowing for the investigation of typical phenomena in high-energy physics. To facilitate their dissemination, the data are shared in accessible and commonly used formats. Additionally, they are accompanied by software and web interfaces designed for easy use, without the need for installation or coding by the user. The objective is twofold. On the one hand, the goal is to promote these activities in outreach and educational contexts, such as Summer Schools, Masterclasses, university projects, as well as various initiatives within the ATLAS Collaboration itself. On the other hand, high-energy physics research becomes accessible to an interdisciplinary audience, involving experts from other fields to benefit from their expertise, such as machine learning and computer science. This work provides an overview of the ATLAS Open Data resources, with concrete examples of how they can be used to promote scientific education and research in the field of particle physics.

1. – ATLAS Experiment

The Large Hadron Collider (LHC) is the largest and most powerful particle accelerator in the world. Located at the European Organization for Nuclear Research (CERN), it extends over 27 kilometers in an underground ring and hosts numerous experiments, the result of collaboration among thousands of researchers from around the world. Essentially, the LHC accelerates particle beams to speeds close to the speed of light. These beams are then directed to collide at predetermined interaction points, around which complex detection apparatuses, commonly known as detectors, are installed. The goal of these collisions is to recreate conditions similar to those of the primordial universe to study the laws that govern the universe and the fundamental constituents of matter.

(*) IFAE 2024 - "Poster" session

Among the experiments conducted at the LHC, ATLAS (A Toroidal LHC ApparatuS) [1] stands out as one of the most significant and comprehensive. Its detection apparatus consists of a cylindrical block weighing approximately 7000 tons and standing 25 meters tall, extending 46 meters around the interaction point. Inside, there are various modules designed in a complementary manner to understand different properties of the particles passing through them, enabling their tracking and identification. This enormous detector has a general-purpose design, allowing for the exploration of a wide range of physical phenomena, from the in-depth study of the Standard Model of particle physics to the search for new forms of physics, such as supersymmetry, dark matter, and exotic particles.

These phenomena are extremely rare, so studying them requires great technical skill, both in recreating favorable conditions for their observation and in analyzing the resulting measurements. It is necessary to generate an enormous number of collisions to collect sufficient data on the rare phenomena of interest. To give an idea, the amount of raw data produced by the LHC in one year is of the order of tens of zettabytes, equivalent to millions of years of high-definition video. Most of this data is discarded immediately through appropriate selections to retain only the most interesting events, actually reducing the amount of stored data to the still impressive order of hundreds of petabytes per year [2, 3]. In addition to the complexity of the physics involved, the huge amount of saved data implies several challenges, ranging from technology for data acquisition, hardware resources to process and store them, trusted pipelines for data management, to advanced techniques to analyze the collected data and monitor resources functioning [4]. For the ATLAS experiment alone, over the years, a software stack of about 10 gigabytes and over 4 million lines of code has been developed to facilitate a centralized and coherent management of the analysis workflows [5]. All these factors constitute significant access barriers, both for participation in research and for the dissemination of knowledge, limiting its fruition to only the experts working in this field.

The ATLAS Open Data project [6] was created precisely in response to these limitations, with the aim of simplifying the training of new generations of researchers and promoting the dissemination of scientific discoveries to the general public. Thanks to this project, the data collected from the ATLAS experiments are made available to students, teachers, and physics enthusiasts worldwide, facilitating the spread of knowledge and stimulating interest in science.

2. – Open Data

The concept of Open Data refers to the practice of making data available to everyone, without access or usage restrictions. This approach promotes transparency, collaboration, and innovation in scientific research. CERN adheres to an Open Data policy based on the FAIR principles [7, 8]: Findable, Accessible, Interoperable, Reusable. These principles ensure that data are easily findable and accessible, can be used interoperably with other data and tools, and are reusable for further research and applications.

In the context of the ATLAS experiment, Open Data includes the data collected during proton-proton collisions, as well as Monte Carlo simulations used to model various physical processes. These data are made available in standardized formats and are accompanied by detailed documentation, code examples, and tutorials to facilitate their use by researchers, teachers, and physics enthusiasts.

CERN's Open Data policy allows anyone to consult and reuse ATLAS data. This means that the data are not reserved exclusively for members of the ATLAS collaboration

but can be used by anyone interested, including scientists from other disciplines, teachers, students, and the general public. This openness not only increases the transparency of scientific research but also promotes public participation, dissemination of results, and interdisciplinary collaboration.

2˙1. *The importance of Open Data for advancing science.* – Sharing data with the public offers numerous benefits both for society as a whole and for the scientific community.

Firstly, it fosters greater public engagement in scientific research. Making ATLAS data available opens the world of research to the entire society. This increases awareness and understanding of scientific processes and research results, promoting a more widespread and informed scientific culture.

In addition, Open Data can serve as an educational tool to support teachers in student education. They provide valuable resources that educators can use to develop teaching materials and practical activities. This support is particularly suitable for student training, offering them the opportunity to work with real data, closely engage with scientific methods, and develop advanced analytical skills.

Another important goal this initiative contributes to is making the outcomes of public funding tangible. The ATLAS experiment, like most research entities, receives substantial public funding from numerous countries, without which sustaining such a complex organization would not be sustainable. Sharing research data and results with the public transparently shows how these funds are used and what the concrete benefits for society are. This increases trust in the scientific process and the value of investments in research.

Finally, besides fulfilling the social responsibilities inherent in research, the Open Data project can be seen as a strategic investment aimed at reaching a broader scientific community. In the current landscape, where interdisciplinary exchange plays an increasingly crucial role in advancing research, Open Data creates an ideal access point to attract expertise from other sectors from which our field can benefit. Consider, for example, the impact of adopting cutting-edge techniques for analyzing and managing the vast amounts of data produced at CERN and the added value that collaborations with professionals in Machine Learning and Computing could bring.

It becomes clear that Open Data represents a valuable resource for promoting transparency, collaboration, and innovation in scientific research. Making ATLAS data accessible to a broader audience allows for maximizing the impact of the research, training new generations of scientists, and engaging society in the wonder of scientific discovery.

3. – ATLAS Open Data

The data shared by the ATLAS experiment within the Open Data project include high-energy collisions and can be divided into three main collections.

- 1 fb^{-1} at 8 TeV (2012) [9]: two datasets shared both in XML format and as ROOT [10] ntuples, totaling approximately 15 million events. This collection is primarily intended for educational and outreach purposes, prioritizing clarity and ease of use over precision and details;

- 10 fb^{-1} at 13 TeV (2015-2016) [11]: includes both real data and Monte Carlo simulations. These data are also intended to support research activities, allowing the exploration of processes described by the Standard Model, evaluation of systematic uncertainties, and search for signals of new physics (Beyond Standard Model);

- 139 fb^{-1} at 13 TeV (2015-2018) [12, 13]: consists of data simulated using Pythia and Delphes, shared as part of the research competition "HiggsML uncertainty challenge" hosted at NeurIPS 2024. The objective is to give researchers the opportunity to work with ATLAS data to develop Machine Learning methods capable of providing predictions with associated uncertainty.

3˙1. *Overview of ATLAS Open Data initiatives.* – ATLAS uses Open Data to organize numerous educational and research initiatives, involving a wide audience on an international scale. One of the main activities is represented by the Masterclasses, aimed at high school students and organized in collaboration with 225 research institutes. These Masterclasses have so far involved over 13,000 students from more than 60 countries, offering them the opportunity to work directly with ATLAS data and to learn the basics of data analysis in particle physics.

Besides Masterclasses, ATLAS also promotes competitions for researchers, such as TrackML [14], HiggsML [15], and HiggsML Uncertainty [12,13]. These challenges aim to solve complex problems related to the analysis of ATLAS data, attracting the interest of experts in machine learning and computer science. They thus have a dual value, as they constitute interesting research problems both from the point of view of the physics results achieved and from the point of view of the methods developed. As evidence of this, these competitions have been hosted by top-tier venues in the Machine Learning field such as the Kaggle platform [16] and the NeurIPS conference [17], attracting the participation of prominent researchers in these disciplines.

3˙2. *Getting started: software, documentation and computing resources.* – To facilitate access and use of ATLAS Open Data, a wide range of software resources has been made available [18, 19]. These include tools for inspecting histograms directly in the browser, allowing users to view and analyze data without the need to download or install additional software. Moreover, event visualization tools are available, providing a detailed graphical representation of collisions, making it easier to understand and interpret the data.

Another very useful tool is Jupyter notebooks, available in Python and C++, which provide code examples and tutorials to guide users in data analysis. These notebooks can be run both locally and on cloud platforms, using Docker containers, virtual machines, and cloud instances. In this way users can choose the most suitable work environment for their needs, facilitating access to data and computing resources necessary for analysis.

Furthermore, comprehensive online documentation is available [6], including detailed guides, examples, and tutorials to support users in using ATLAS Open Data. This documentation is constantly updated to reflect the latest versions of the data and tools, ensuring that users have access to the most recent and accurate information.

4. – Conclusions

The ATLAS Open Data project is a valuable resource that achieves multiple goals. It engages the general public by making complex scientific data accessible, thereby promoting broader scientific literacy. It nurtures the next generation of researchers through educational resources and simplified datasets tailored for students and educators. Additionally, it fosters interdisciplinary collaborations, bringing together diverse expertise to advance scientific progress. By democratizing access to high-energy physics data, the project supports ongoing research and cultivates a well-informed, skilled scientific community.

REFERENCES

[1] ATLAS Collaboration, *J. Instrum.*, **3** (2008) S08003.
[2] Clissa Luca, Lassnig Mario and Rinaldi Lorenzo, *Front. Big Data*, **6** (2023) 1271639.
[3] Clissa Luca, *Supporting Scientific Research Through Machine and Deep Learning: Fluorescence Microscopy and Operational Intelligence Use Cases*, PhD Thesis (2022) http://amsdottorato.unibo.it/10016/.
[4] Clissa Luca, Lassnig Mario and Rinaldi Lorenzo, *Comput. Softw. Big Sci.*, **6** (2022) 16.
[5] Krasznahorkay Attila, *Developing Education and Outreach Resources From Research Data*, https://indico.jlab.org/event/459/contributions/11675/ (accessed 08/07/2024).
[6] ATLAS Collaboration, *Atlas Open Data Website*, https://opendata.atlas.cern/ (accessed 08/07/2024).
[7] CERN, *Open Data Policy for the LHC Experiments*, Technical Report, CERN-OPEN-2020-013, Geneva (2020).
[8] Wilkinson Mark D. *et al.*, *Sci. Data*, **3** (2016) 1.
[9] ATLAS Collaboration, *Dataset 8 TeV*, https://opendata.atlas.cern/docs/documentation/overview_data/data_education_2016/ (accessed 08/07/2024).
[10] Antcheva Ilka *et al.*, *Comput. Phys. Commun.*, **180** (2009) 2499.
[11] ATLAS Collaboration, *Dataset 13 TeV*, https://opendata.atlas.cern/docs/documentation/overview_data/data_research_2024 (accessed 08/07/2024).
[12] ATLAS Collaboration, *HiggsML Uncertainty Challenge (Codabench)*, https://www.codabench.org/competitions/2164/ (accessed 08/07/2024).
[13] ATLAS Collaboration, *HiggsML Uncertainty Challenge (NeurIPS)*, https://fair-universe.lbl.gov/ (accessed 08/07/2024).
[14] ATLAS Collaboration, *TrackML Particle Tracking Challenge (Kaggle)*, https://www.kaggle.com/competitions/trackml-particle-identification (accessed 08/07/2024).
[15] ATLAS Collaboration, *Higgs Boson Machine Learning Challenge*, https://www.kaggle.com/c/higgs-boson (accessed 08/07/2024).
[16] *Kaggle*, https://www.kaggle.com.
[17] *Conference on Neural Information Processing Systems*, https://neurips.cc/.
[18] ATLAS Collaboration, *Atlas Software for 2016 Open Data Release*, http://doi.org/10.7483/OPENDATA.ATLAS.WS1A.RLES (2016).
[19] Lo Ya-Feng, *ATLAS Open Data: Software for Visualization and Physics Analysis*, https://cds.cern.ch/record/2655357 (2019).