ATL-SOFT-PROC-2025-013
20 January 2025

Open Data at ATLAS: Bringing TeV collisions to the World

Giovanni Guerrieri[1,*], on behalf of the ATLAS Collaboration

[1]CERN

Abstract. ATLAS Open Data for Education delivers proton–proton collision data from the ATLAS experiment at CERN to the public along with open-access resources for education and outreach. To date ATLAS has released a substantial amount of data from 8 TeV and 13 TeV collisions in an easily-accessible format and supported by dedicated documentation, software, and tutorials to ensure that everyone can access and exploit the data for different educational objectives. Along with datasets, ATLAS also provides data visualisation tools and interactive web based applications for studying the data, along with Jupyter Notebooks and downloadable code enabling users to further analyse data for known and unknown physics cases. The Open Data educational platform which hosts the data and tools is used by tens of thousands of students worldwide, and we present the project development, lessons learnt, impacts, and future goals.

1 Introduction

Open data constitutes one of the most effective means for scientific collaborations to disseminate knowledge and establish a long lasting legacy. It serves a dual purpose, not only enabling and advancing research outside of the lifetime of the experiments, but also fostering education by providing students and educators with authentic datasets to explore and learn from.

The ATLAS Open Data initiative [1] covers both education and research efforts. The education project, which is the focus of these proceedings, offers open access to proton–proton collision data gathered by the ATLAS experiment [2] at the Large Hadron Collider (LHC) [3]. The provided resources are developed collaboratively with students and educators, and aim at high school, undergraduate, and postgraduate learners, as well as individuals who are not formally associated with any institution. On the other hand, the recent Open Data for research release [4] comes with sufficient detail to be used for new scientific publications.

The project implements the FAIR principles (Findable, Accessible, Interoperable, Reusable) [5], by providing unique identifiers to the resources, adopting standardised protocols for data management, and supporting cross-platform interoperability. These practices integrate well with ATLAS Open Data's core values of accessibility, usability, and transferable expertise. The resources are designed to be inclusive, enabling a wide range of users, regardless of technical background or location, to achieve diverse learning objectives. Beyond High Energy Physics (HEP), participants gain valuable skills in software development

[*] e-mail: giovanni.guerrieri@cern.ch

and machine learning. The open datasets and related tools are widely utilised in classrooms, public events, masterclasses, and international workshops. An overview of the past and future Open Data releases for education is presented in Section 2. Section 3 describes the tools and the software available for users, whereas the infrastructure and platforms on which the software operates are covered in Section 4. In Section 5 the best practices to ensure usability in Open Data are discussed. Section 6 summarises the overall ATLAS Open Data activity and describes the future plans.

2 The ATLAS Open Data releases for Education

The ATLAS Open Data releases feature a collection of datasets gathered during the ATLAS detector's data acquisition runs[1] at the LHC. These datasets include 8 TeV and 13 TeV proton–proton collision data, with both real collision events and Monte Carlo simulations, along with dataset variations to estimate systematic uncertainties.

The datasets are curated with calibrated and simplified information about reconstructed physics objects, making them accessible and manageable for a wide range of users. The inclusion of the "*for education*" label in plot titles such as the one shown in Fig. 1 ensures clarity of purpose, distinguishing these datasets as tools tailored for learning and exploration.

The releases include:

- **8 TeV (2016)** [6, 7]: this release features 1 fb^{-1} of LHC Run 1 data, representing approximately 4.5% of the 2012 dataset (~6 GB).
- **13 TeV (2020)** [8]: datasets released as part of the LHC Run 2 data-taking, featuring 10 fb^{-1} of data, equivalent to about 30% of the 2016 dataset (~150 GB).
- **13 TeV (2025)**: The upcoming release will offer 36 fb^{-1} of data, further enhancing the scope for educational analysis. The data will be provided as a flat ROOT `NTuple` [9].

2.1 Resources location and preservation

Each release is accompanied by datasets hosted on the CERN Open Data Portal [10] and documentation, tutorials, and supporting resources available on the ATLAS Open Data website [1]. These materials guide learners in analysing particle-physics data in educational environments, offering a hands-on experience that emulates real-world research processes.

All data are shared under a Creative Commons CC0 waiver [11], with unique DOI identifiers assigned to facilitate proper citation in scientific works. The ATLAS-approved publications mentioned above detail the content, properties, and recommended usage of the datasets. While simplified for educational purposes, the datasets retain complexity and require an undergraduate-level understanding of particle physics.

[1] The term "Run" is used to describe a specific period of data taking, which can last for several years, in which both the experiments and the LHC are operated under reasonably stable conditions.

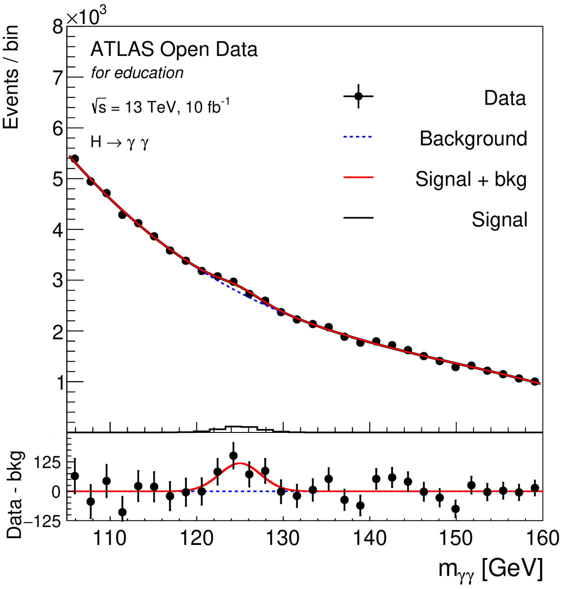

Figure 1. Example of analysis performed with the 2020 release for education. Figure found in Ref. [8].

3 Software and Tools

To complement the open datasets, the ATLAS collaboration provides a suite of software and tools [12] to enable users to analyse and interpret HEP data. Among these tools is the Histogram Analyser, a web-based platform designed to quickly and intuitively interact with cut-based data analysis. This tool allows users to visualise datasets through interactive histograms, focusing on how physicists distinguish between different physics processes. By adjusting the range of data for different variables, users can isolate specific signals, such as Higgs boson events, from background processes. This hands-on approach is particularly useful for building expertise in interpreting data distributions and optimising selection criteria for advanced analyses.

Accompanying the Histogram Analyser is a comprehensive collection of Jupyter notebooks [13] designed for interactive data analysis. These notebooks cover a wide range of topics, from Standard Model (SM) analyses to Beyond the Standard Model (BSM) physics.

The notebooks are meant to be accessible directly through a web browser, and target users with with varying levels of expertise. They also support multiple programming frameworks, including Python, C++, RDataFrame [14], and Uproot/Coffea [15, 16], ensuring that users can work in a flexible environment and have examples in the environment with which they are most familiar. Beyond physics applications, these resources include tools for training and statistical analyses that extend to other fields, such as data science and related curricula. This versatility makes the ATLAS Open Data resources valuable for broader educational purposes.

A series of YouTube tutorials [17] is also provided, offering step-by-step guidance on using the tools, conducting analyses, and understanding the datasets.

4 Computing infrastructure

Extensive support for computational resources and infrastructure is provided, to accommodate a variety of use cases and user expertise levels, offering web-based (online) resources, and hybrid solutions for limited internet access, that also allow fully offline functionality. The main components of the infrastructure are summarised below, highlighting their capabilities and suitability for different scenarios.

4.1 Web-based interactive analysis

CERN's SWAN (*Service for Web-based Analysis*) [18], the **ESCAPE's VRE** (*Virtual Research Environment*) [19], and **Binder** [20], are three platforms offering web-based environments for interactive data analysis. They all leverage Jupyter notebook technologies, enabling users to write and execute code while visualising results inline.

SWAN is a CERN-developed platform that combines advanced computing infrastructure with seamless integration into CERN services, such as CERNBox [21] for data storage. Supporting Python, C++, and ROOT, SWAN provides the tools necessary for analyses. Its integration with CERN's infrastructure ensures access to robust computing capabilities and powerful data management resources, such as Rucio [22]. However, SWAN requires CERN credentials, which limits its accessibility to external users.

The **Virtual Research Environment (VRE)** addresses this limitation by providing a platform similar to SWAN but accessible to non-CERN users. It enables users to create accounts without CERN credentials, by using the INDIGO IAM token issuer. By bridging the accessibility gap, the VRE extends the reach of web-based data analysis to a broader audience while maintaining the relevant computational capabilities.

In contrast, **Binder** provides a lightweight alternative independent of authentication services, making it globally accessible without the need for credentials. Users supply a GitHub repository containing their notebooks, and Binder generates a Docker image to execute the analysis, all at the click of a button. Its accessible nature and reusable links foster collaboration, though its computational resources are more limited than those of SWAN or the VRE.

4.2 Docker

Docker resources [23] enhance reproducibility and ease of setup for ATLAS Open Data analyses. The project provides pre-configured Docker images, distributed via the GitHub container registry associated with the ATLAS Open Data repository. Docker's containerisation ensures that analyses can be run consistently across different systems, resolving common compatibility issues.

Docker is particularly advantageous for collaborative and educational settings, where maintaining consistency and minimising setup complexity are crucial. By shipping all dependencies and settings in a container, users can easily share analytical environments locally or on analysis facilities such as SWAN.

4.3 Virtual Machines

The ATLAS Open Data project provides **Virtual Machines (VMs)** as an offline option for analysing datasets. These VMs replicate a virtual operating system on the user's machine, preloaded with the necessary 13 TeV software and tools. They offer a viable, low-effort solution for offline analysis, specifically for educational use or environments with unreliable internet access.

4.4 Main Features Across Platforms

By offering a diverse set of tools, the ATLAS Open Data infrastructure ensures that users have access to the resources they need to perform analyses, whether online, offline, or in collaborative environments. The main advantages of these resources include:

- **Accessibility:** Tools like Binder prioritise global accessibility, while SWAN and Docker offer high-performance options for users with specific credentials or advanced needs.
- **Reproducibility:** Docker and VMs ensure modular, consistent environments for robust and reproducible analyses.
- **Flexibility:** The listed platforms support various coding languages, as well as accommodating different environments and analysis setups.
- **Collaboration:** Analysis facilities such as SWAN or the VRE enable collaborative work by supporting the sharing of data, software, and computational environments among users.

5 Ensuring the Usability of Open Data: Key Practices and Challenges

Several factors can significantly impact the usability of Open Data initiatives. Data may become inaccessible if its location changes or if it gets corrupted without proper tracking or backup systems in place. Outdated or deprecated analysis tools and dependencies can potentially limit the effectiveness of the tools. Insufficient or outdated documentation can lead to confusion, making it harder for users to navigate the data and tools. Access restrictions, due to permission settings or infrastructure limitations, can also create barriers, preventing users from using Open Data resources. Furthermore, without sufficient maintenance, infrastructure can become prone to errors, inconsistencies, and prolonged downtime, which disrupts user workflows and data accessibility.

To mitigate these challenges and ensure long-term usability, several best practices should be implemented and maintained throughout the lifecycle of Open Data initiatives:

- **Manage Code in Versioned Repositories**: storing code in version-controlled repositories, such as Git, ensures that changes are tracked and previous versions can be accessed. This practice allows users to work with reliable, reproducible and consistent code.
- **Package the Analysis Environment in Software Containers**: containers, such as Docker, encapsulate the analysis environment, including dependencies, tools, and code, ensuring that the environment is reproducible across different systems. This minimises configuration inconsistencies and guarantees that analyses can be performed consistently, regardless of the underlying infrastructure.
- **Document Everything from the Start**: comprehensive and up-to-date documentation is essential for usability, but it goes beyond providing instructions. Integrating documentation with issue tracking allows issues to be systematically recorded and addressed. This approach helps maintain clarity and alignment between tools, methods, workflows, and user feedback.
- **Define Easily Reusable Workflows**: workflows should be modular and designed with reuse in mind. While notebooks simplify this process, it remains essential to adopt a structured approach that emphasizes consistency. Standardising the organisation, documentation, and execution of workflows ensures they are straightforward to adapt and share across different projects or teams.

- **Use Continuous, Automated Testing**: implementing continuous integration and automated testing ensures that updates to the codebase do not introduce errors or break existing functionality. Automated tests help verify the correctness of analysis workflows, and enable timely identification of problems before they impact users.

By following these best practices, Open Data efforts can significantly reduce the likelihood of issues that hinder usability, ensuring a more efficient, user-friendly environment. However, it is recommended to also remain committed to vigilance, which is a crucial requirement in order to tackle unforeseen challenges.

6 Conclusions

Open Data initiatives in ATLAS have been serving educational purposes since 2016, providing computing resources, tutorials, and documentation that support the datasets. These assets are widely used by several institutions for training sessions, workshops, and masterclasses, fostering the growth of HEP-based learning across disciplines.

A new 13 TeV Open Data for education release is coming soon, offering opportunities to adopt new data formats and expand the available dataset to higher integrated luminosities. There are also new possibilities emerging from the recent Open Data for research release. It is crucial to identify synergies and complementarities between the education and research efforts. This not only includes adopting shared practices in providing documentation and resources, but also planning for events such as workshops and hackathons; a common ATLAS education + research Open Data event is currently in the organisation phase.

Monitoring and evaluation efforts are crucial to secure a bottom-up, community-based approach towards activities. Alongside usage statistics and user surveys, an ongoing priority is to keep improving the documentation. Data alone has a limited scope; an equal value resides in the users' ability to access, understand, and use resources effectively.

Finally, the Open Data community is expanding, with new examples and contributions being collected from around the world. This collaborative approach ensures that the ATLAS Open Data initiative continues to evolve and serve a wide range of needs and use cases.

References

[1] *ATLAS Open Data website*, https://opendata.atlas.cern

[2] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, *JINST* **3** (2008), S08003, https://doi.org/10.1088/1748-0221/3/08/S08003

[3] L. Evans, P. Bryant, LHC Machine, *JINST* **3** (2008), S08001, https://doi.org/10.1088/1748-0221/3/08/S08001

[4] ATLAS Collaboration, `DAOD_PHYSLITE` format 2015-2016 Open Data for Research from the ATLAS experiment, CERN Open Data Portal (2024), http://doi.org/10.7483/OPENDATA.ATLAS.9HK7.P5SI

[5] M. Wilkinson, M. Dumontier, I. Aalbersberg, *et al.*, The FAIR Guiding Principles for scientific data management and stewardship, *Sci Data* **3** (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[6] ATLAS Collaboration, Review of the ATLAS Open Data Dataset, *ATL-OREACH-PUB-2016-001* (2016), https://cds.cern.ch/record/2203649/

[7] ATLAS Collaboration, Review of ATLAS Open Data 8 TeV datasets, tools and activities, *ATL-OREACH-PUB-2018-001* (2018), https://cds.cern.ch/record/2624572

[8] ATLAS Collaboration, Review of the 13 TeV ATLAS Open Data release, *ATL-OREACH-PUB-2020-001* (2020), https://cds.cern.ch/record/2707171

[9] R. Brun and F. Rademakers, ROOT – An Object Oriented Data Analysis Framework, *Nucl. Inst. & Meth. in Phys. Res. A* **389** (1997) 81-86, https://doi.org/10.5281/zenodo.848818

[10] *CERN Open Data portal*, https://opendata.cern.ch

[11] C. Commons, CC0 1.0 Universal License, https://creativecommons.org/publicdomain/zero/1.0/

[12] *ATLAS Outreach data and tools repository*, https://github.com/atlas-outreach-data-tools

[13] *Project Jupyter*, https://jupyter.org/

[14] D. Piparo , *et al.*, RDataFrame: Easy Parallel ROOT Analysis at 100 Threads, *EPJ Web of Conferences* **214** (2019), 06029, https://doi.org/10.1051/epjconf/201921406029

[15] J. Pivarski , *et al.*, scikit-hep/uproot3: 3.14.4, *Zenodo* (2024), https://doi.org/10.5281/zenodo.4537826

[16] N. Smith , *et al.*, Coffea Columnar Object Framework For Effective Analysis, *EPJ Web of Conferences* **245** (2020), 06012, https://doi.org/10.1051/epjconf/202024506012

[17] *The ATLAS Exepriment YouTube channel*, https://www.youtube.com/playlist?list=PL1qU3k-RDRsvy3jhxUTmq7ZJQTdFvjLJn

[18] D. Piparo, E. Tejedor, *et al.*, SWAN: A service for interactive analysis in the cloud, *Future Generation Computer Systems* **78** (2018), 1071-1078, https://doi.org/10.1016/j.future.2016.11.035

[19] E. Gazzarrini, E. Garcia, *et al.*, The Virtual Research Environment: towards a comprehensive analysis platform, *2305.10166* (2023), https://arxiv.org/abs/2305.10166

[20] *The Binder Project*, https://mybinder.org/

[21] H. Gonzalez Labrador, *et al.*, CERNBox: Storage gateway for CERN and beyond, *EPJ Conferences* **295** (2024), http://dx.doi.org/10.1051/epjconf/202429501012

[22] M. Barisits, *et al.*, Rucio: Scientific Data Management, *Comput. Soft. Big Sci.* **3** 11 (2019), https://doi.org/10.1007/s41781-019-0026-3

[23] *The Docker website*, https://www.docker.com/