The First Release of ATLAS Open Data for Research

Zachary Marshall[1,*], *Lukas Alexander* Heinrich[2], *Mario* Lassnig[3], *Mariana Isabel* Vivas Albornoz[4], and *Stephane* Willocq[4] on behalf of the ATLAS Computing Activity

[1]Lawrence Berkeley National Laboratory
[2]Technische Universität München
[3]CERN
[4]University of Massachusetts

Abstract. The ATLAS Collaboration has released an extensive volume of Open Data for Research use for the first time. The full datasets of proton collisions from 2015 and 2016, alongside a wide array of matching simulated data, are all offered in the PHYSLITE format. This lightweight format is chosen for its efficiency and is the preferred standard for ATLAS internal analyses. Additionally, the inclusion of Heavy Ion collision data considerably widens the scope for research within the particle physics community. To ensure accessibility and usability, the release includes a comprehensive suite of software tools and detailed documentation, catering to a varied audience. Code examples, from basic Jupyter notebooks to more complex C++ analysis packages, aim to facilitate engagement with the data. This contribution details the available data, corresponding metadata, software, and documentation, and initial interactions with researchers outside the ATLAS collaboration, underscoring the project's potential to foster new research and collaborations.

ATL-SOFT-PROC-2025-007
11 January 2025

1 Introduction

Open Data, along with scientific publications, form the foundation of the legacy of modern scientific experiments. Publications are critical, but they are static: they cannot change to account for new understanding of theoretical calculations, or to implement new ideas for the discovery of new particles. Open Data can provide the flexibility required to carry scientific value long after the lifetime of the experiment itself. Because these experiments are paid for by public funding, Open Data also ensure that the research outputs are accessible more broadly and preserved indefinitely.

Four different "levels" of Open Data were recognized by the Data Preservation for HEP working group [1]. The simplest are additional supporting data for a publication: digitized tables and figures, or statistical likelihoods, for example. ATLAS provides ample supporting data via the HEPData repository [2], as well as via specific web pages for each publication. The second level is simplified data, like those used for outreach and education. ATLAS has released significant Open Data of this sort, along with documentation, that have been used in tutorials and educational efforts around the world [3]. The third level is Open Data with

*e-mail: ZLMarshall@lbl.gov

sufficient detail to be used for new scientific publications. While previously, ATLAS had provided some variety of bespoke datasets for specific applications that were of this level of detail, only in July 2024 were the first general-purpose Open Data for Research released by the collaboration. These Open Data, their support and documentation, and a wide variety of associated issues are the focus of these proceedings. The fourth level of Open Data is the raw detector data. In the case of ATLAS, these are too complex, and require too much processing to be of value outside the experiment. They will be preserved internally, but there is no intention to release them.

Together with the other LHC experiments, ATLAS agreed to regularly release Open Data for Research [4]. 25% of the data from each Run[1] is to be released five years after the Run, and an additional 25% is to follow five years later. The data are to be released in a format that is used for internal data analysis. They also must come with sufficient Monte Carlo (MC) simulation and documentation to be useful to researchers outside the experiment. The new release of Open Data for Research is the first step along this path for ATLAS.

The subsequent sections describe several aspects of the Open Data for Research release. The data themselves are explained in Section 2, and the accompanying software is described in Section 3. The documentation that has been prepared for the release is described in Section 4. The resources available for working with the Open Data are described in Section 5, and the support model for the Open Data by the collaboration is described in Section 6. Monitoring the usage of the Open Data is important to understanding its value to and adoption in the community, and is described in Section 7. Finally, plans and considerations for the future are described in Section 8.

2 The Release

On July 1, 2024, ATLAS released the 2015 and 2016 proton–proton collision data at a center of mass energy of 13 TeV, totaling 7.1 billion events and 45 TB, along with about 300 MC simulation datasets totaling 2 billion events and 20 TB [5]. All these samples have been made available on the CERN Open Data portal [6], and significant documentation is provided on the ATLAS Open Data website [3]. The samples are all released under the Creative Commons CC0 license [7] with a request for citation of the dataset and collaboration when they are used. Because "data" cannot be protected by copyright, the CC-BY license that requires attribution cannot be applied.

Each MC simulation dataset corresponds to a different physics process (e.g. di-jet production, or Z-boson production). The nominal samples used by most ATLAS analyses are included for almost all Standard Model processes, alongside a variety of variation samples (e.g. with different event generators) for the determination of systematic uncertainties. For each of these datasets, considerable metadata has been collected and provided, including the sample cross section, filter efficiency (e.g. in case a lepton filter is applied during event generation), k-factor (i.e. correction to a higher-order theoretical calculation of the cross section for the sample), number of events and sum of event weights for proper normalization of the samples, process name and integer identifier, generator software versions, sample keywords, a human-readable description of the physics process, and a link to the exact sample definition in the software.

The data have all been provided in PHYSLITE format [8], a light-weight analysis format that is also used internally for data analysis. Because it is used for published data analysis, this format is rather complicated and includes a great deal of information that might be foreign

[1] In this context, a Run is a several-year long data-taking campaign wherein the detector and accelerator complex are in a reasonably constant configuration.

to new users; therefore, documentation of all variables in the format has been provided on the web with human-readable variable definitions and types. Tools are under development to provide convenient access to the metadata and branch documentation via a pip-installable package.

The data were released via the CERN Open Data portal, and announcements were made in a wide variety of places including the ATLAS public webpage, the CERN Open Data portal, the CERN EP Newsletter, and various social media channels (e.g. Facebook, YouTube, and X). The decision was made early on to post minimal information in all of these places and instead point all users to the ATLAS Open Data website [3]. This helped ensure that any incorrect information could be centrally updated in one place without the need to interact with or correct many different sites or deployments. It also ensures a single point of entry for the primary documentation and for users who want to ask for help.

3 The Software

The standard ATLAS software, called Athena, is publicly available [9] under the Apache 2.0 license [10]. For each stable release, new containers are built for the `x86` and `aarch64` platforms and are distributed via CERN's container registry [11] and via CernVM-FS [12]. The documentation, including tutorials for software development and for analysis use-cases, are also publicly available [13].

Most users are expected to create smaller ROOT ntuples or HDF5 files from the PHYSLITE data that are released, since operating directly on 65 TB of data is often unwieldy, or at least inefficient. An example of creating such an ntuple is provided as a part of the standard ATLAS analysis tutorial [13], and the software used to create the Open Data for outreach and education will soon be released as an additional example.

Once these smaller data formats are in hand, it is expected that users will take advantage of the extensive, mostly python-based data analysis software. Python notebooks have been provided as examples, with significant in-line documentation to ensure that they can stand alone. These notebooks can be run in services like SWAN at CERN or Binder (with container configurations provided for Binder directly in the notebook repository), so that users don't need any local software configurations to get started. Notebooks are provided based on several different python analysis tool ecosystems, including ROOT-based and ROOT-free versions.

For convenience, a python script has also been provided for converting the PHYSLITE-formatted data into json files that can be used by the Phoenix event display [14]. Along with a number of example events that are provided on the web, this allows users to quickly visualize their favorite events or explore the ATLAS detector geometry directly in their web browser. These tools have proven very useful for short tutorials and visual learners.

While in principle this provides everything a new user needs to be able to work with the ATLAS software, create new MC simulation samples, and develop new analysis code, in practice a number of features harm the usability for the general public. For example, the general analysis tutorial explains how to run event generators in Athena. ATLAS currently distributes over 100 GB of parton distribution functions (PDFs) via CernVM-FS?, and packing those into an image would make it almost unusable. As a temporary measure, the minimal set required for the tutorial are included, which results in a container image that is still well over 10 GB.

Similarly, many systems depend on the availability of a CERN login for convenience or for authentication. The ATLAS software is available on the CERN GitLab instance, which allows the use of CERN's authentication and identity mechanisms (e.g. egroups) for managing access permissions. Unfortunately, GitLab at CERN does not allow the general public

to create an account or interact via issues or merge requests. Using GitHub would allow more interaction with non-members, at the expense of more complicated and manual access management that could significantly affect the operation of the experiment. For this reason, the central software is provided only on GitLab, but the notebooks and some of the smaller educational projects that have been developed are available on GitHub.

4 The Documentation

Inspired by the Turing Way [15], the goal was set to provide multiple learning pathways that would be approachable to users with a variety of backgrounds. Additionally, because significant documentation already exists for Open Data for education and outreach purposes, and to minimize the required effort, a secondary goal of minimizing duplication between the paths and learners was set. For example, only one introduction to particle physics needs to be written, and that introduction can be skipped in the case that the person identifies themselves as a theoretical physicist. A number of stand-alone modules have also been built up for specific use-cases like statistics, machine learning, or geometry exploration through event displays.

The immediate goal is to create a continuous stream of material from rather basic, fast lessons (like the event display material) to rather complex material that could take a week or more to work through, with sufficient clarity to allow a user to be injected anywhere along the path depending on their learning goals and knowledge. One key way this has been addressed is by making the creation of the light-weight ROOT ntuples for outreach and education one of the central examples of use of the Open Data for Research. That way, users creating their own ntuples can follow the outreach and education tutorials for use of the ntuples in notebooks, for example, without the need for an additional downstream course module.

The eventual goal is to create an interconnected web of modules that is forward- and backward-navigable. That way, if someone attempts a downstream module (like the machine learning module), they will be able to easily move upstream to attempt more complex or time-consuming material, through the outreach and education tutorials, and even into the research Open Data tutorials as they learn more.

In many cases, simple python notebooks have been provided and suffice for quick lessons. These are excellent for inline code documentation, minimal start-up time and overhead, and having extensive documentation available on the web. For opening a file and making a quick plot, or for analyzing small ntuples that have been made for outreach and education, they are ideal.

For more complex and complete analyses, users are pointed directly to the official ATLAS analysis tutorial. This has the advantage that there is a single, fairly well-maintained tutorial presented to all users who wish to work with the Open Data for Research. The disadvantage is that much of ATLAS's analysis ecosystem is C++-based, so there is an interoperability problem and a jump in difficulty when a user wishes to make their own ntuples for a specific use-case. One additional difficulty is that while the standard ATLAS analysis tutorial will continue to be updated to the latest releases, the tutorial for the Open Data will need to stay consistent with the Open Data themselves. Breaking changes are not frequent, but as they occur the older version of the tutorial has to be lifted out and set aside to be retained for users of the Open Data.

In many cases, it is clear that the additional documentation is also a great asset to ATLAS users. Thanks to the Open Data for Research, more clear and extensive documentation is available for the dataset nomenclature rules, for example. Similarly, there are more accessible explanations of systematic uncertainties, their origins, their types, and their use in analysis. Much of this documentation is evergreen, and an attempt has been made to split the evergreen

portion that will not change from the technical documentation that depends on the exact version of software being used, for example.

Because the data are in PHYSLITE format, there are also limitations to what research can be done with them. For example, PHYSLITE does not contain all tracks, so users cannot perform track-counting analyses. Similarly, particle flow objects are not provided, so users cannot build their own jets to test new jet algorithms or grooming algorithms. Only limited searches for long-lived particles can be performed with the PHYSLITE data format. The documentation of what is possible, and of the limitations, is important in order to ensure that users do not spend significant time going down a dead-end, or draw incorrect physical conclusions from searches or measurements that are inaccurate because of the limitations of the data format itself.

5 The Resources

The data occupy some 65 TB of disk space and are currently provided through the Open Data Portal at CERN. CERN IT have agreed to provide sufficient storage resources for the LHC experiments' Open Data for this initial period, which is a great advantage to the experiments who do not request resources for Open Data as a part of their standard resource requests. These resources are not guaranteed for all time: it is not clear whether this will be considered a "host laboratory responsibility", if these resources will continue to be available for the next several years, or if they will be provided even beyond the lifetimes of the experiments. The data have been placed on a Rucio [16] storage endpoint at CERN, so that they can be quickly and easily transferred to any other institute with such an endpoint. For users who have CERN accounts, the standard CERN resources (batch system, notebook platforms like SWAN, etc.) are all available and can be used to access and process the data.

Of course, this presents an immediate equity concern. Only users with quite significant institutional resources (both in terms of storage and in terms of network access) can afford to pull down 65 TB of data for analysis, and can afford the CPU to process those data. It is quite important to provide alternative options to those with limited resources and without CERN connections alternative access mechanisms.

To that end, documentation has been provided for the use of standard commercial cloud resources like Google Cloud and AWS. These systems provide significant free, well-connected computing, and the data can be streamed from the CERN Open Data portal so that they do not need to be transferred into the cloud. For larger users, more hours can be purchased, and the cost for hosting data in the cloud is not significant as long as the amount of data being extracted is relatively limited. The option of storing a copy of the data in various commercial cloud providers is also being explored.

Several sites have expressed an interest in providing notebook-like access to users with basic authentication (e.g. via Google or GitHub accounts), and the University of Nebraska at Lincoln has already provided such a site [17]. There are some concerns around abuse and security that have prevented a more widespread adoption of this solution to date, but with the growth of federated identity programs and better sand-boxing of notebook platforms and containers, it might be possible to spread this solution in the future.

6 The Support

Mechanisms for support are also critical, as users' questions need to be answered, but the collaboration cannot be burdened with a significant support load. The CERN Open Data Forum [18] offers an entry point for all users, with a category specifically for ATLAS Open

Data support and various tags to identify specific problems. A mailing list has also been constructed for ATLAS members who are willing to provide support for Open Data. All of these are best-effort mechanisms: no funding mechanism or additional effort has been identified for the support of Open Data to date. Documentation and support have to be balanced carefully in order to maximize the return on the minimal effort available. Edge cases that only one or two users might ever stumble across might not be worth the effort, additional complexity, and detail of additional documentation (one does not read an encyclopedia cover-to-cover, and indexing becomes a major issue when documentation is sufficiently extensive). Instead, it might be better to reserve the effort for directly responding to users in the forums in case anyone does stumble upon those issues. Because the CERN Open Data Forum is built on Discourse, answers are preserved and reasonably well indexed so that future users can benefit from any answers provided there.

While the websites and documentation point users to these support forums, it is clear that users also reach out for help through any channel available to them. Shortly after the release of the Open Data for Research, a user asked for help via Reddit. With postings and announcements across social media, there are a wide variety of places where users might comment or ask for more information that need to be monitored, and ATLAS members have been asked to be on the lookout for lost users who need to be redirected to more appropriate support channels. This makes it even more critical that users are pushed to the central ATLAS Open Data website as early on in their interaction as possible.

An open question is the level of support that will be offered to researchers wishing to publish new, unique papers. In some cases these might require additional MC simulation samples to be released, or additional statistics for existing samples; this has been foreseen and a straightforward approval mechanism for the release of additional MC simulation has been created. There is no current mechanism for the review of a paper from an external author by the collaboration prior to publication or standard journal review. There are always concerns, of course, about the risk of a person publishing a paper with an extraordinary claim that would require a response from the collaboration. As the collaboration gains experience with these Open Data, the full picture and effort required to effectively support research should become more clear.

For users that are sufficiently advanced, want deeper support, or want access to additional data, the collaboration offers a Short Term Associate [19] mechanism. This allows external members to join the collaboration for targeted projects that can culminate in research papers, and provides full access to ATLAS internal information during the period of association. One of the hopes in releasing the Open Data for Research is that some users will be drawn to this mechanism after seeing that the data are reasonably approachable.

7 The Usage

Monitoring the usage of the Open Data is critical to understanding their adoption and success. There are a number of metrics that can be automatically gathered and tracked with minimal effort. These include forks and stars of GitLab and GitHub repositories. The Open Data Portal records have DOIs that can be cited in academic works, and the citation counts are monitored and reported in the portal itself. Experience shows that these are not always diligently cited, and some effort is ongoing to automatically monitor papers posted to arXiv to identify possible uses of the ATLAS Open Data without citations. For the data themselves, the Open Data Portal can track downloads and clicks. Some storage monitoring is available to understand accesses as well, for example for users at CERN who directly access the data or for users of cloud platforms that stream the data instead of downloading them. Within the ATLAS Open Data website, there is monitoring of the search terms that users enter, so that

documentation can be provided or improved that specifically targets areas where users seem to be having difficulty.

In addition to these automatic metrics, the ATLAS Outreach and Education team has run periodic surveys of Open Data users to understand usage and satisfaction with various aspects of the Open Data (the data themselves, their documentation, and so on). These surveys are key to get better narrative explanations of what works and doesn't. Of course, whenever an event is held (e.g. a tutorial or workshop), it is possible to quickly gather feedback; often these events feature a rush to improve documentation immediately prior, and a second rush to fix problems that were identified immediately afterwards.

One key issue is that a number of educational use-cases are not easily tracked by any of these metrics. A university group might download or locally store (a subset of) the data, avoiding the download tracking. They might build their own notebook-based courses for working with the data — indeed, this has happened at a number of universities worldwide with the existing Open Data. They are unlikely to publish academic papers, so citations are not identified or tracked for these projects. Finding metrics that track these users effectively is quite important, and this is an ongoing area of discussion. As a starting point, the courses and projects that have already been developed are being gathered on the ATLAS Open Data website, so that the community is aware and able to take advantage of existing work.

One of the future targets for monitoring is the use of CPU for processing the Open Data. While most of the above metrics track access, there is currently very little understanding of how much CPU is used (at CERN or worldwide) for processing Open Data, running notebook-based analyses, tutorials, etc. Gathering these data will be challenging, as identifying which users are processing Open Data is not trivial, and the data will need to be aggregated from a variety of sources. In some cases, like the commercial cloud resources, it is likely to prove impossible to fully monitor what processing users are doing. Nevertheless, this could provide valuable information about the resources required to practically support the use of Open Data for Research that is not available today.

8 The Future

This release is only the first of many to come for the experiment. It also provides an excellent opportunity to understand and consider how the experiment will evolve over the next 30 years, and what issues might arise:

- Who has ownership of the code and data, and will ownership automatically transfer in case a person leaves the experiment or retires?
- Where are the data stored, and is the storage sufficiently resilient to not require intervention in the future? This applies not just to the 'primary' Open Data for Research, but to small derived datasets or files that might be used in examples on the websites or in tutorials.
- How is the website hosted, where is the code stored, and who maintains it? Is it sufficiently well constructed to be ported to new hosting systems? What if one of the underlying tools (e.g. Docusaurus [20]) is replaced in 10 or 20 years?
- How is the software run? Virtual machines were the path forward 10 years ago; today Docker containers are the standard. What will be the standard in 20 years?
- Is the internal documentation sufficiently good to allow easy extension of the Open Data, for example to add a new MC simulation sample or additional statistics in the same format as the existing samples? If they are deleted, can they be reproduced?
- The data formats will evolve in the coming years; should the Open Data be entirely re-released, or should the newer Open Data be allowed to exist in a format different from the

older Open Data? If it is re-released, should the old data format be deleted, or retained and supported? Can all the documentation be updated if the format is changed? What about users who rely on the old format and don't wish to update their own software or notebooks?

These are not just theoretical issues; almost all of them have impacted work on Open Data within ATLAS over the last few years. As the collaboration ages, it is likely that many of them will also impact the internal data formats, analysis tools, and documentation as well. Not only do these Open Data offer an opportunity to examine how things might evolve over the life of the experiment, they also offer a baseline for long-term data preservation within the experiment. That is, if at the end of the lifetime of the experiment, there are not sufficient effort and resources available to continue to support more complex internal analysis data formats, the Open Data for Research will stand as the permanent analysis data even for internal usage. The issue of long-term data preservation has already affected the experiment, as the Run 1 data (data collected prior to 2013) are largely unsupported, and discussions of the long-term support for Run 2 and Run 3 data (data collected up to 2026) are ongoing.

At the end of 2024, ATLAS released the next batch of Open Data for Research, including for the first time heavy ion collisions. A fresh release of Open Data for outreach and education is also coming soon, with more extended and updated analysis examples and tutorials in development. An Open Data workshop is also in the planning phases, to enhance the documentation and provide users with ample opportunity to try things out and see for themselves how the Open Data can be used, with guides from the experiment to get them started. Finally, efforts are continuing to gather examples of good Open Data use from around the world. Several examples have already been identified from courses on at least four continents, and as the use spreads around the globe the ATLAS Open Data website will offer an opportunity for users to share their projects, learn from one another, and take part in the ATLAS Open Data community.

References

[1] Z. Akopov et al., *Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics*, DPHEP-2012-001 (2012), 1205.4667

[2] *HEPData*, https://www.hepdata.net

[3] ATLAS Collaboration, *ATLAS Open Data* (2023), http://opendata.atlas.cern

[4] *CERN Open Data Policy for LHC experiments*, CERN-OPEN-2020-013 (2020), https://cds.cern.ch/record/2745133

[5] ATLAS Collaboration, CERN Open Data Portal (2024), doi:10.7483/OPENDATA.ATLAS.9HK7.P5SI

[6] *CERN Open Data Portal* (2024), http://opendata.cern.ch

[7] Creative Commons, *CC0 1.0 Universal License*, https://creativecommons.org/publicdomain/zero/1.0/

[8] ATLAS Collaboration, *Software and computing for Run 3 of the ATLAS experiment at the LHC*, CERN-EP-2024-100 (2024), 2404.06335

[9] ATLAS Collaboration, *Athena Git Repository* (2023), http://gitlab.cern.ch/atlas/athena

[10] *Apache Licence, Version 2.0* (2004), https://www.apache.org/licenses/LICENSE-2.0

[11] *CERN Harbor Registry*, https://registry.cern.ch/harbor

[12] *CVMFS*, https://cvmfs.readthedocs.io/en/stable/

[13] *The Athena Framework*, https://atlassoftwaredocs.web.cern.ch/athena/athena-intro/
[14] *Phoenix event display*, https://github.com/HSF/phoenix
[15] *The Turing Way*, https://book.the-turing-way.org/index.html
[16] M. Barisits et al., Comput. Soft. Big Sci. **3**, 11 (2019)
[17] *Opendata Analysis Facility T2_US_Nebraska*, https://coffea-opendata.casa/hub/login
[18] *CERN Open Data forum*, https://opendata-forum.cern.ch
[19] ATLAS Collaboration, *Collaborating with ATLAS*, https://atlas.cern/Discover/Collaboration/External-Collaboration
[20] *Docusaurus*, https://docusaurus.io