ATLAS Data Access Policy

ATLAS Collaboration, September 2014 (revised: February 2021)

Introduction

ATLAS has fully supported the principle of open access in its publication policy. This document outlines the policy of ATLAS as regards open access to data at different levels as described in the DPHEP [1] model. The main objective is to make the data available in a usable way to people external to the ATLAS collaboration.

In 2020, the LHC Open Data Working Group (ODWG) released a policy document for open data [2], which extends the level-3 open data policy described below.

The ATLAS policy for data preservation is described in a separate document. The collaboration's need to preserve data for its own use shares some requirements with making them open access. To support open access to data additional resources will be required to develop and support the tools to make the data available.

Policies for Different Data Levels

Open access to ATLAS data by people outside the collaboration can be considered at four levels of increasing complexity, listed below, with associated conditions, see Ref. [1]. This policy pertains to collision physics data (i.e. that are stored offline and intended for physics analysis) and the necessary associated metadata, along with associated simulated datasets and tools allowing to produce new simulated datasets based on an adequate simulation of the ATLAS detector.

The sharing of data with non-ATLAS members for purely technical studies (e.g. file reading technicalities or new data formats), and not for evaluating their physics content, may be approved by ATLAS on a case-by-case basis. In these cases the data should not be shared beyond those involved in the approved collaboration.

Level-1. Published results

All scientific output is published in journals, and preliminary results are made available in Conference Notes. All are openly available, without restriction on use by external parties beyond copyright law and the standard conditions agreed by CERN.

Data associated with journal publications are also made available: tables and data from plots (e.g. cross section values, likelihood profiles, selection efficiencies, cross section limits, …) are stored in appropriate repositories such as HepData [3]. ATLAS also strives to make additional material related to the paper available that allows a reinterpretation of the data in the context of new theoretical models. For example, an extended encapsulation of the analysis is often provided for measurements in the framework of RIVET [4]. For searches information on signal acceptances is also made available to allow reinterpretation of these searches in the context of models developed by theorists after the publication. ATLAS is also exploring how to provide the capability for reinterpretation of searches in the future via a service such as RECAST [5].

RECAST allows theorists to evaluate the sensitivity of a published analysis to a new model they have developed by submitting their model to ATLAS.

Level-2. Simplified data for outreach & education, and collaboration with non-HEP sciences

Outreach and education

ATLAS recognizes the vital role of outreach and education, participates in and encourages outreach and education activities, and makes selected data available for them. Typically, a fraction of the complete ATLAS data-set is used, selected to provide a rich sample of events with interesting physics signatures but not adequate for a publication of a physics result. The data are provided in simplified, portable and self-contained formats for educational and public understanding purposes. Lightweight environments to allow the easy exploration of these data are also provided.

This includes event displays of selected events, ntuple or similar level data for illustrating the calculation of invariant mass distributions, lifetimes, CP asymmetries, etc. ATLAS collision data or simulated data or both may be used. ATLAS has already made such data public for a diverse set of purposes, ranging from analysis programs and event displays for high-school and university education to releasing selected subsets of simulated data events for Data Science related contests [6]. ATLAS level-2 data would also be suitable as an input to more universal, multi-experiment outreach interfaces. Examples of the full complexity of the analysis data and environment may also be provided for educational purposes.

The data are provided for educational and public understanding purposes only, and may not be used for publication of physics results.

Collaboration with non-HEP sciences

ATLAS encourages scientific collaboration with other science communities, e.g. computer science or education. This collaboration may use specific ATLAS datasets (simulated events or collision data) and/or ATLAS tools. Projects where ATLAS datasets are used should proceed via the established association programme (STA, MCI). Upon completion of the scientific collaboration project, the dataset used for the project and the tools required to use/access it may be made public for unrestricted use, if they are not sensitive and after confirmation by the ATLAS spokesperson.

Level-3. Reconstructed data

ATLAS recognizes the potential value of making its reconstructed (level-3) data openly available after reasonable embargo periods. Any release of the data must be accompanied by a release of appropriate simulated datasets and tools to perform a data analysis. The ODWG document [2] describes the LHC-wide open level-3 data policy.

By far the most practical means of conducting a specific new analysis that requires level-3 data is in partnership with the collaboration. In particular, the ATLAS collaboration has established association programmes for specific analyses. Non-ATLAS members can submit a proposal, supported by an ATLAS member, to collaborate with ATLAS on a particular topic of interest. These proposals are considered on a case-by-case basis and if granted give access to the ATLAS data and the internal information required for this analysis.

Level-4. Raw data

It is not practically possible to make the full raw dataset from an experiment of the scale of ATLAS usable in a meaningful way outside the collaboration. This is due to the complexity of the data, metadata and software, the required knowledge of the detector itself and the methods of reconstruction, the extensive computing resources necessary and the access issues for the enormous volume of data stored in archival media. It should be noted that, for these reasons, general direct access to the raw data is not even available to individuals within the collaboration, and that instead the production of reconstructed data (as discussed above) is performed centrally. Therefore, ATLAS does not propose to devote resources to developing access to the full raw dataset for non-members of the collaboration during the active lifetime of the collaboration. Access to representative smaller samples of raw data might be considered if well motivated.

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*", arXiv:1205.4667 [hep-ex]

[2] "*CERN Open Data Policy for the LHC Experiments*", ATL-COM-CBPOLICY-2020-003, https://cds.cern.ch/record/2747308

[3] http://hepdata.cedar.ac.uk/

[4] RIVET: http://rivet.hepforge.org/

[5] Kyle Cranmer and Itay Yavin, "*RECAST: Extending the Impact of Existing Analyses*", JHEP 1104:038, 2011

[6] www.kaggle.com/c/Higgs-boson