ATLAS Data Preservation

RWL Jones1, DM South2, KS Cranmer3 on behalf of the ATLAS Collaboration

1 Department of Physics, Lancaster University, Lancaster LA1 4EJ UK
2 DESY, Hamburg and Zeuthen, Germany
3 Department of Physics, New York University, New York, New York, USA

E-mail: Roger.Jones@cern.ch

Abstract. Complementary to parallel open access and analysis preservation initiatives, ATLAS is taking steps to ensure that the data taken by the experiment during Run-1 remain accessible and available for future analysis by the collaboration. An evaluation of what is required to achieve this is underway, examining the ATLAS data production chain to establish the effort required and potential problems. Several alternatives are explored, but the favoured solution is to bring the Run 1 data and software in line with the equivalent to that which will be used for Run 2. This will result in a coherent ATLAS dataset for the data already taken and that to come in the future.

1. Data Preservation, Data Access and Data Sharing

Data preservation is a heavily overloaded term. ATLAS [1] is careful to distinguish data preservation – the ability to reanalyze data or derived data at a later date – from the issues of data sharing. Further, it sees distinct use cases for internal reuse, external data sharing while the collaboration is still active and reuse after the collaboration has ceased to be active. While all aspects have their merits, they have different priority, and while the experiment is still active the preservation for internal reuse is a high priority. This use case can often produce tools and formats of use for the other use cases.

Even the data sharing during the active life of the experiment can be broken into important sub-cases: principally data sharing for education and outreach; and data sharing for research performed by third parties. ATLAS has always regarded outreach and education as an important part of its activity. In terms of data sharing for additional research, ATLAS regards the scientific utility as an important consideration, and accordingly places its resources into tools that have most likely robust scientific outcome.

ATLAS has also attempted to learn from the experiences of earlier experiments (at DESY, SLAC and FERMILAB), perhaps best summarized in the DPHEP report [2]. One commonly held lesson is that the aim is more accurately described as information preservation, as the data alone has little meaning without a software environment and full documentation, which must also include the tacit information. It is important also to triage the data preserved. While it is technically feasible to archive every bit produced beyond the trigger system, this would be both wasteful and prohibitively expensive. It is also important to plan the data preservation from an early stage in the experiment. This has therefore been one of the considerations in the recent redesign of the ATLAS event data model [3].
Accordingly, ATLAS has produced several documents, including internal documents on the planning for data preservation and a mandate to formulate recommendations for analysis preservation policy and implementation, and externally visible documents [4] describing the data preservation policy, data access policy.

2. ATLAS Data Preservation at Various Levels
ATLAS has adopted the DPHEP decomposition of data preservation into 4 levels in its planning.

2.1. Publication, Education and Outreach Data
The highest level is the additional documentation associated with papers. From the start, ATLAS has made good use of HEPDATA [5] and HEP-inspire [6] for this purpose, and fully scrutinized additional information is encouraged for each analysis. Equally, ATLAS has had an active education and outreach activity since first data taking. Several tools allow interaction with real ATLAS experiment data, although not for research purposes [7]. This activity has lead to the release of over 2 fb⁻¹ of Higgs data (in 4 lepton and 2 photon modes) in ATLAS analysis-level formats, as well as other real data samples of general interest. This is a very active area of development, and ATLAS has recently formed a ATLAS outreach data and tools working group. The existing tools and samples are now linked via the CERN Open Data Portal [8]. An interesting addition to the outreach area is the ATLAS Kaggle Higgs Challenge [9], encouraging research into improved machine learning approaches to the Higgs search. This includes large simulated data sets that could be repurposed for other studies.

2.2. Analysis and Raw Data
In the DPHEP analysis, level 3 preservation requires the analysis level data and environment be preserved, while level 4 requires enough of the raw data be preserved such that full reprocessing remains possible. ATLAS requires both levels of preservation for internal purposes, although access to the reprocessing data and environment is limited even within the collaboration, and will not in general be possible for third parties; the long-term preservation does not require the ability for full reprocessing once the experiment is inactive. Preservation at levels 3 and 4 presents a choice. One can either:

1. make a final processing of the data with a fixed software/environment and maintain the latter forever;
2. or periodically reprocess with new software.

The latter option retains all possibilities to benefit from new algorithms, improved calibration techniques etc, which is important for internal reuse. The first option requires back-porting of new algorithms, matching of new calibrations to old formats and code to gain full benefit from future use. As ATLAS wishes to be able to make use of old data in future analyses, it regards the additional load of reprocessing with new software to be less than that of back-porting to earlier environments. It also avoids several technology issues. This choice introduces several requirements.

2.2.1. Real Data Processing, Reprocessing and Analysis. It must be possible to read, process and analyse real data with the desired conditions and new software versions. This in turn means backwards compatibility must be retained for the format to be processed, even if new detectors are added. This poses particular challenges for rapidly changing items, such as trigger objects. It also means the reconstruction software must work in an optimal and meaningful way for old data. In order to allow processing and analysis, a ‘best knowledge’ (BK) tag of the associated conditions needs to be preserved for each period of running. This tag must then be migrated through all future technologies. Any software requiring new conditions must be able to derive them from the old conditions, or else use a dummy; similarly, all downstream conditions must be derivable in an automated way.
2.2.2. Simulation and digitization. There are also important requirements on the simulation and digitization. Simulation of newly generated Monte Carlo must be possible with the same software and conditions matching any real data. For this, the first requirement is that new Geant [10] versions must be verified as describing the old detector well enough. Similarly, fast simulation must continue to give a good description of older data.

The simulated data must be digitizable, taking into account the correct readout, pile-up, beam conditions etc to match the real data. The trigger simulation is particularly problematic, as it depends on a fixed offline software version at the time of data taking, so in this one case old software must be used. This digitized data must then be reconstructable with the same software as the real data; this must reflect time dependent effects such as the radiation damage to detectors.

2.2.3. Metadata. All the above, plus the needs of analysis, place the requirement that all appropriate metadata (including the LHC beam conditions, the ATLAS data taking and quality conditions etc) must remain available corresponding to the real data. Detector status information is a special case, as this may change with software version, as new features in the software may require detector information that renders some of the data unusable.

2.3. Analysis Preservation

Analysis preservation is an important, but less settled area. Even if just for documentation, all analyses need to be captured, and reuse of the analysis tools developed is very useful for future analyses. These objectives do not place the strict requirement that an analysis be exactly replicated, such that with the same data, conditions and software the same results should be obtained at the bit level.

2.3.1. Reproducibility. The strict condition referred to above is sometimes termed reproducibility. Such strict reproducibility may be a suitable case for the use of capture using virtual machines. ATLAS is assessing the level of requirement for this reproducibility. There are open questions about the expected lifetime for virtual machines, the degree of separation from the hardware layer that is possible and the ability to capture every nuance of an analysis, including the tacit information.

2.3.2. Replicability. The more general repetition of an analysis using new data, tools, conditions etc is sometimes referred to as replicability. This is a more generally useful approach, but requires a high degree of forward porting; however, it is aligned with the general ATLAS data preservation strategy. However, there is likely to be a finite lifetime within which even this level of replicability is possible; it is likely that the overhead of porting beyond a major change in the event data model is not practicable, and re-implementation of algorithms when needed is a better approach beyond that point. ATLAS will be ensuring replicability of the Run 1 data through the next run. However, after a major breakpoint in the environment evolution, recasting solutions [11] (where the final output of an analysis is captured in terms of a generalised phase space, and may then be compared with new physics model generators, which are run though a form of the analysis selections) become appropriate. While the software is being forward ported, the combined performance of the analysis object tools must be redetermined, which in practice means this must be possible either automatically or with almost no effort.

2.3.3. Validation. Both reproducibility and replicability require a robust validation system. ATLAS has several of these optimized for different aspects of the online and offline processing changes. These are being assessed as part of the analysis preservation task force, and it is likely these will be combined into an automated validation system dedicated for data and analysis preservation.

3. Conclusions
ATLAS is committed to the long-term preservation of its data for scientific use and reuse, within the limits of available resource. Accordingly, it is directing its activities to areas where the evidence of scientific benefit is clearest, or else the use for education and outreach is clear.

A detailed plan for general data preservation, primarily driven by the requirements for internal reuse has been produced and is being implemented. The work has potential benefits even for the current running period. The requirements and plans for analysis preservation is the subject of an ongoing task force that is due to report in summer 2015. In general, ATLAS efforts focus on the forward-porting of software to allow old data to be used with new software versions. For use by third parties, ATLAS is particularly exploring the potential of recasting of analyses as a means of providing a robust mechanism for new physics models to be tested against well-validated analysis chains by non-members of the collaboration.

References
[1] The ATLAS Collaboration “The ATLAS Experiment at the CERN Large Hadron Collider” JINST 3 S08003
[2] DPHEP Status Report 2012 arxiv:1205.4667.
[3] J. Catmore (2015) “New Petabyte-scale Data Derivation Framework for ATLAS”, Proceedings of the 2015 Int. Conf. on Computing in High Energy and Nuclear Physics 2015, to be published in J. Phys. Conf. Ser.
[4] The ATLAS Collaboration (2015) “ATLAS Data Preservation policy” ATL-CB-PUB-2015-002 http://cds.cern.ch/record/2012333/files/ATL-CB-PUB-2015-002.pdf ; The ATLAS Collaboration (2015) “ATLAS Data Access policy” ATL-CB-PUB-2015-001 http://cds.cern.ch/record/2002139/files/ATL-CB-PUB-2015-001.pdf.
[5] A Buckley and M Whalley (2010) “HepData reloaded: reinventing the HEP data archive”, Proceedings of 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, PoS ACAT2010:067
[6] https://inspirehep.net/info/general/project/index
[7] M. Pedersen, F. Ould-Saada, M.K. Bugge (2015) “Sharing ATLAS data and research with young students”, Procs ICHEP 2014, to be published in Nucl. Phys. B Proc. Supp, ATLAS-OREACH-PROC-2015-001 https://cds.cern.ch/record/1984338/files/ATL-OREACH-PROC-2015-001.pdf
[8] http://opendata.cern.ch/education/ATLAS
[9] https://www.kaggle.com/c/higgs-boson
[10] S Agostinelli et al. (2003) “GEANT4: a simulation toolkit” NIM A596 250-303; J Allison et al. (2006) “GEANT4 developments and applications”, IEEE Transactions on Nuclear Science 53 No. 1 (2006) 270-278
[11] K Cranmer and I Yavin “RECAST: Extending the Impact of Existing Analyses” JHEP 1104:038,2011