Analysis Preservation in ATLAS

Kyle Cranmer[1], Lukas Heinrich[1], Roger Jones[2], David M. South[3], for the ATLAS collaboration

[1]New York University, New York, USA
[2]Lancaster University, Lancester, UK
[3]Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany

E-mail: `lukas.heinrich@cern.ch`

Abstract. Long before data taking, ATLAS established a policy that all analyses need to be preserved. In the initial data-taking period, this has been achieved by various tools and techniques. ATLAS is now reviewing the analysis preservation with the aim of bringing coherence and robustness to the process and with a clearer view of the level of reproducibility that is reasonably achievable. The secondary aim is to reduce the load on the analysts. Once complete, this will serve for our internal preservation needs but also provide a basis for any subsequent sharing of analysis results with external parties.

1. Introduction

The ATLAS experiment [1] is one of four major experiments at the Large Hadron Collider (LHC) [2] at CERN. During the first run of data-taking from 2009 to 2012 (Run 1) 8 TeV proton-proton collision data corresponding to an integrated luminosity of about $20\,\mathrm{fb}^{-1}$ has been collected and analyzed. The accumulated data for Run 1 across all experiments reached around 100 PB during Run 1. With experiments of this scale, which are not easily repeatable, it is imperative to have a robust strategy for ensuring the reproducibility and accessibility of the scientific results of these experiments. Therefore, not only must there be a strategy for preserving the data but also for the procedures used to analyze them. The ATLAS collaboration has been committed to a comprehensive strategy to preserve both data and analyses and efforts are underway to deepen and expand the scope of this preservation.

2. Analysis Preservation during Run 1

2.1. Dataset Provenance Information

The raw data as recorded by the detector is centrally reconstructed using the ATLAS reconstruction software framework *Athena*. The first pass reconstruction is done using a fixed release ensuring that all real and simulated data is processed using mutually compatible software — a policy referred to as *frozen Tier0* [3]. Changes to the software, such as performance enhancements, were only allowed if the reconstruction output was not affected by it. Functional improvements which alter the reconstruction output, due to increased understanding of the software and data, were implemented via full reprocessings of the entire dataset. This approach ensured that a coherent representation of the real and simulated data was available at all times.

During Run 1, most ATLAS analyses relied on derivative datasets produced from the output of the reconstruction software framework. Most importantly, these derivative datasets were natively processable by the widely used data analysis framework ROOT [4]. Additionally, the content of these derived datasets, which were produced for each physics working group (e.g. Top, SUSY, Exotics, etc.) separately, were tailored for the group that produced them and are documented internally.

Which exact software release was used and how the reconstruction was performed for the data used in a given analysis is crucial information. This information is stored in the Atlas Metadata Interface (AMI) [5], which can be queried both programmatically as well as using a web interface.

2.2. Combined Performance Recommendations

During Run 1, the information regarding ATLAS analyses was preserved in a number of ways. Traditionally, many analyses are prepared to be published at a handful of large community conferences organized over the course of a year such as the *Recontres de Moriond* conference held in early spring. For these conferences, recommendations for data analysis procedures that reflect the state-of-the-art knowledge about the data are disseminated throughout the collaboration. Recommendations can include, among other things, physics object definitions, prescriptions on how to estimate backgrounds or recommendations for the usage of specific triggers and are prepared by the combined performance and physics working groups. These recommendations are preserved in both easily accessible and quickly modifiable collaborative web services such as wikis but also in more permanent internal notes once the recommendations are finalized.

Together with recommendations, the combined performance groups release software tools that aid with implementing them. A typical analysis needs a sizable number of these packages (e.g. 20-50) which must be compatible with each other. On top of the documentation and archiving of the individual packages via version control software and internally published instructions, many physics working groups compiled curated lists of such packages and sometimes reference implementations of the recommendations.

2.3. Analysis Software and Documentation

Datasets and combined performance recommendations are shared among many analyses. The more specialized information regarding the implementation of a concrete analysis has been documented mainly through the detailed supporting documents that are prepared during the course of an analysis. These documents, that are accessible internally, describe not only analysis techniques and additional studies that have been performed but often also technical information on how the analysis code was run. In addition, it has been ATLAS policy to require all analysis code to be checked into version control to preserve it for later reference.

2.4. Open Access and Additional Analysis Data

In addition to preserving the details of an analysis internally, the collaboration is committed to providing additional material and data related to an analysis to the wider high energy physics community or even the general public. Where possible, information such as signal acceptances or distributions that have been corrected for detector effects (i.e. *unfolded*) have been published to the online HEP results repository HepData [6]. The information stored there can be then analyzed using non-collaboration software packages such as Rivet [7]. Additionally, the collaboration has regularly released specifically prepared material including versions of the recorded data for both collaboration with fields outside of particle physics, such as the machine learning community via the *Higgs Machine Learning Challenge* [8], or outreach and educational purposes [9, 10].

3. Plans for Run 2

While the efforts described in the previous section already capture a great number of details regarding the analyses of the ATLAS data, there are many opportunities to improve on them. Despite the available documentation, it is in practice often quite involved to trace exactly how a given result was produced. The necessary information is scattered over many different repositories, such as the metadata interface, the various source code repositories, internal documents and web-pages. Additionally, frequently a number of even smaller derivative datasets, containing the minimum amount of necessary information, are produced privately by the analyzers (i.e. not organized by the collaboration) to improve the performance of the data analysis. Therefore, often the only people that can realistically reproduce the results are those that were part of the original analysis team. This poses a potential problem for long-term preservation, in the case that people take on different responsibilities in the collaboration or even leave the field entirely. Therefore, ATLAS is committed to increase the coherence and robustness of the analysis preservation efforts based on the experience gained in Run 1. The different aspects described below are also being coordinated with various community-wide initiatives such as DASPOS [11] and DPHEP [12].

3.1. Reproducibility vs. Replicability

For analysis preservation ATLAS has identified two general paradigms to study, each with different scope and advantages and termed *reproducibility* and *replicability.*[1]

Reproducibility describes the concept of archiving existing software, tools and documentation regarding the used analysis procedures. In this sense an analysis is reproducable if one can in detail redo the steps the original analysis team has undertaken to produce the results. For this to be possible, all ingredients that went into the results need to be preserved in the state they were in at the time of publication. This includes computer configuration (e.g. operating system and architecture), software releases in use at the time, and the datasets as they were then reconstructed. These requirements make it mostly useful for short and medium term preservation. Reproducibility is therefore most applicable for confirmation and clarification of the published result, e.g. in the case that a faulty analysis procedure has been identified and the collaboration wants to verify whether the published result is affected by it. Additionally, it is useful for applications regarding reinterpretation of the analysis with respect to alternative signal models which is described in more detail in Section 3.3.

For the implementation of the reproducibility paradigm, various options have been considered. An attractive method of preservation is via the use of virtualization techniques. With virtualization, it is possible to preserve exact software and hardware configurations without relying on physical availability of outdated systems. ATLAS is investigating technologies such as full virtual machine (VM) images, but also more limited but flexible containerization solutions such as Docker [13] or Linux Containers (LXC) [14]. Containers are a more lightweight example, and likely the most useful for preservation of individual data analysis applications.

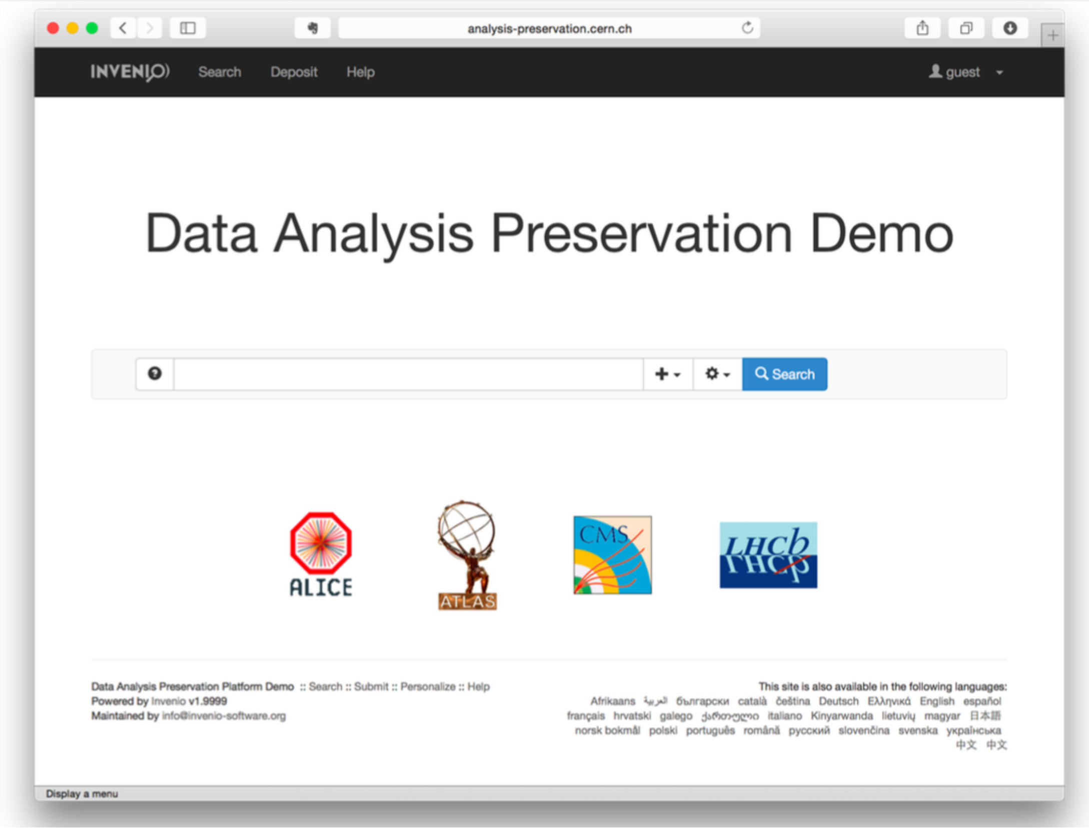
Conversely, replicability refers to the process of ensuring that analyses leading to published results are repeatable using the most recent version of software tools and data formats. In particular, given the enormous amounts of data, it will not be possible to keep datasets reconstructed with old software releases on storage indefinitely. Therefore, it is important to keep the old data readable by newer versions of the software. Regarding specific analyses, ATLAS is investigating options to ensure replicability in this sense. This will most likely be achievable via code migration and regression testing as well as detailed human-readable information on how

[1] As both concepts described in this section ultimately aim to *reproduce* the analysis results based on the preserved information, it has been challenging to find appropriate terminology, which unambiguously differentiates the two. For clarity and consistency with previous documents we use replicability and reproducibility while conceding linguistic inelegance.

the analysis has been performed. Use cases for this approach are, among others, extending the analysis using additional data collected after the original publication, or conservation of know-how across the collaboration. Newer members of the collaboration will not be familiar with older software and analysis procedures, so relying solely on reproducibility, as defined above, is not sufficient.

3.2. Centralized Preservation of Analysis Metadata

Together with the other three LHC experiments and the CERN IT department an effort is underway to capture all relevant analysis metadata in a centralized web-accessible location [15]. This will facilitate the access to information which is currently scattered around different services for each experiment. The system tries to capture data from all stages of the analysis, beginning with the original datasets, the analysis code, various metadata information on all used signal and background samples, but also documentation, discussion and related publications such as conference presentations. An early prototype is in preparation, a screenshot of which is presented in Figure 1.



Data Analysis Preservation Demo

Data Analysis Preservation Platform Demo :: Search :: Submit :: Personalize :: Help
Powered by Invenio v1.9999
Maintained by info@invenio-software.org

This site is also available in the following languages:
Afrikaans العربية български català čeština Deutsch Ελληνικά English español
français hrvatski galego ქართული italiano Kinyarwanda lietuvių magyar 日本語
norsk bokmål polski português română русский slovenčina svenska українська
中文 中文

Figure 1: The Data Analysis Preservation web interface.

3.3. Reinterpretations of Preserved Analyses

An important use-case and incentive for preserving analyses is the opportunity to reinterpret existing analyses in light of new physics signals that for various reasons could not be considered in the original publication. While many analyses are optimized to look for specific models of physics beyond the Standard Model, it is likely that the same analysis is actually quiet sensitive to a broader range of such models. While analysis teams try to incorporate such additional signal hypotheses in their searches, often due to external constraints this cannot be done comprehensively. For example, new theoretical candidate models might become known only after publication or the analysis team might lack resources to carry out additional studies.

As described above, currently it is rather hard for collaborators outside of the original team to then reimplement the analysis. Therefore, ATLAS is actively prototyping and evaluating a framework, called RECAST [16, 17], that will make use of the above preservation efforts (relying on the concept of *reproducibility*) to provide a streamlined interface that will allow the wider high energy physics community to suggest reinterpretations and the collaboration to effectively honor such requests.

4. Conclusion

The ATLAS experiment is committed to a comprehensive preservation of the analyses that lead to the published results in order to ensure reproducibility of these results. The many ingredients that are part of such an analysis have been preserved in different formats during Run 1. In light of the new round of data-taking, ATLAS is revisiting the analysis preservation procedures to improve the fidelity of the captured information. Together with the other LHC experiments it aims to store relevant meta-data in a centralized repository. Also it is investigating possibilities for the capture of analysis code and configurations for the short and long term using virtualization techniques (for *reproducibility*) or migration (*replicability*) to ensure that as much information regarding existing analyses is preserved as is realistically possible, ideally ensuring that they can be re-run after publication. In addition to archival purposes, this information will be useful for the reinterpretation of results in the context of new physics models.

References

[1] ATLAS Collaboration 2008 *JINST* **3** S08003

[2] Evans L and Bryant P 2008 *Journal of Instrumentation* **3** S08001

[3] Seuster R 2012 *Journal of Physics: Conference Series* **396** 022043

[4] Brun R and Rademakers F 1997 *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **389** 81 – 86 ISSN 0168-9002 new Computing Techniques in Physics Research V

[5] Albrand S, Doherty T, Fulachier J and Lambert F 2008 *Journal of Physics: Conference Series* **119** 072003

[6] The Durham HepData Project URL `http://hepdata.cedar.ac.uk/`

[7] Buckley A, Butterworth J, Lonnblad L, Grellscheid D, Hoeth H *et al.* 2013 *Comput.Phys.Commun.* **184** 2803–2819 (*Preprint* `1003.0694`)

[8] Cowan G, Rousseau D and Bourdarios C 2015 URL `http://cds.cern.ch/record/2007301`

[9] Pedersen M, Ould-Saada F and Bugge M K 2015 Sharing ATLAS data and research with young students Tech. Rep. ATL-OREACH-PROC-2015-001 CERN Geneva URL `http://cds.cern.ch/record/1984338`

[10] ATLAS OpenData Portal URL `http://opendata.cern.ch/education/ATLAS`

[11] Data and Software Preservation for Open Science URL `http://daspos.crc.nd.edu/`

[12] Data Preservation and Long Term Analysis in High Energy Physics URL `http://www.dphep.org/`

[13] Docker URL `http://www.docker.com/`

[14] Linux Containers URL `http://linuxcontainers.org/`

[15] Data Analysis Preservation Demo URL `http://analysis-preservation.cern.ch/`

[16] Recast Control Center Demo URL `http://recast-demo.cern.ch`

[17] Cranmer K and Yavin I 2011 *Journal of High Energy Physics* **2011** 38