

Analysis Preservation and Systematic Reinterpretation within the ATLAS experiment

Kyle Cranmer, Lukas Heinrich on behalf of the ATLAS collaboration

New York University 726 Broadway, New York, NY 10003, USA

E-mail: lukas.heinrich@cern.ch

Abstract. The LHC data analysis software used in order to derive and publish experimental results is an important asset that is necessary to preserve in order to fully exploit the scientific potential of a given measurement. An important use-case is the re-usability of the analysis procedure in the context of new scientific studies such as the reinterpretation of searches for new physics in terms of signal models that not studied in the original publication (RECAST). We present the usage of the graph-based workflow description language *yadage* to drive the reinterpretation of preserved HEP analyses. The analysis software is preserved using Docker containers, while the workflow structure is preserved using plain JSON documents. This allows the re-execution of complex analysis workflows on modern distributed container orchestration systems and enables a systematic reinterpretation service based on such preserved analysis.

1. Introduction

The discovery of the Higgs boson by the ATLAS and CMS collaborations[1, 2, 3] during first data-taking period (‘Run-1’) of the Large Hadron Collider (LHC) has completed the Standard Model of Particle Physics (SM). Besides the continued precision measurements of the Standard Model parameters such as the Higgs boson mass, the search for phenomena predicted by Beyond Standard Model (BSM) theories is a major focus of the LHC collaborations. While there is a large number of plausible BSM scenarios, constraints on both human and computational resources in the collaborations prohibit dedicated data analyses for each proposed extension of the Standard Model. Hence, the collaborations focus on a selection of data analyses targeting complementary final states whose design is guided either by concrete BSM scenarios or so-called “simplified models”[4] acting as proxies for a class of phenomenologically similar theories. No significant excess has been observed thus far in these analyses and typically results are presented as statistical inference results on the parameters of the guiding (simplified) model.

These published limits are not easily translatable into constraints on BSM theories that are not well-aligned with the assumptions of the original BSM model. However, often the analyses are still sensitive to such alternative models, even though they were not optimized for them. This provides an opportunity to still derive limits by *re-interpreting* (or *re-casting*) the analysis in a two-stage procedure:

- (i) Generation of simulated events of the new candidate theory, including reconstruction and detector simulation (“Signal Generation”)
- (ii) A re-execution of the analysis (including event selection and statistical evaluation) of new signal paired with archived background estimates and observed collision data. (“Analysis”)



The appeal of this approach is that it allows a broad survey of possible models by efficiently re-using the limited set of analyses. Notably such a reinterpretation is much less resource-intensive than a dedicated analysis for the new candidate theory: It only requires the relatively light-weight simulation of a new signal, while keeping the resource-intensive analysis procedure design as well as the data and SM background estimates fixed.

As most BSM searches are not unfolded, the above recipe can only be accurately implemented by the LHC collaboration itself as both the detector simulation (using Geant4[5]), the analysis pipeline (using the collaboration-specific Event Data Model and analysis software) as well as necessary data inputs are not readily accessible for non-collaboration members. While third-party implementations exist for both signal generation (e.g. via Delphes[6]) and analysis execution via a number of analysis catalogs provided by Rivet[7] and CheckMate [8], those can only provide approximations that rely on published performance and analysis-specific data, such as reconstruction efficiency tables and resolutions. As analyses adopt complex multivariate event selections or non-standard reconstruction objects, such re-implementations may not be feasible anymore. With the well-structured processes inside of the ATLAS collaboration to generate new signals, the barrier to reinterpretation has been in preserving the analysis pipeline in a reusable way.

2. Analysis Preservation via Declarative Workflows and Linux Containers

2.1. Continuous Software Preservation with Container Images

High energy physics analyses are characterized by their highly distributed development model, where individual processing stages such as event selection and the statistical evaluation are provided by sub-teams within the collaboration mostly without centralized knowledge of how to execute the entire analysis pipeline. The preservation problem thus factorized across into two sub-problems:

- (i) preserving software and computing environment for individual processing stages
- (ii) preserving the workflow logic how to assemble those stages into a coherent analysis pipeline

For the former, the recent advances in Linux Container technology from large industry vendors has enabled a realistic framework of capturing the highly diverse code bases used within an analysis — from systematically developed analysis and event processing frameworks to individual one-off shell scripts — in a standardized, archivable, and portable software environment.

A collaboration-wide effort is now underway to provide base container images adhering to the Open Container Initiative (OCI) image specification [9] along two broad paths: on the one hand as “lean containers” providing the base operating system that are then paired with domain specific high-energy physics (HEP) software taken from the global read-only filesystem CVMFS[10] and on the other hand as “standalone” containers that are independent from an external filesystem. These images are then distributed via a container image registry (such as Docker Hub).

With an increased usage of modern hosted version control services (such as GitHub and GitLab) and their integration with continuous integration services, those images then facilitate the continuous software preservation of individual analysis code that are built into analysis-specific (or sub-stage specific) container images — ensuring that preservation happens in a automated manner *during* the development of the analysis, avoiding the need for post-facto preservation that may be difficult to achieve.

2.2. Declarative Container-based Workflows

Once the individual software environment is preserved, the workflow can be conveniently described in the declarative workflow specification *yadage*. The specification is defined as pure JSON schemas for long-term archivability and native interoperability with the digital library

software Invenio[11] that is widely used within HEP. The language, described in more detail in reference [12], provides a extensible framework for describing arbitrary, run-time dependent directed acyclic graphs (DAGs) of individual container workloads. Within that framework, individual processing stages are defined as “packaged activities” (packtivities) parametrized such that they can be executed on variable inputs that are provided as a JSON document (see Figure 2), while the workflow DAG is modeled as a number of *stages* that extend a directed acyclic graph by scheduling one or more instances of packtivities (graph nodes) and defining their execution dependencies (graph edges) as soon as all required inputs are known.

2.3. Workflow Archival in the CERN Analysis Preservation Portal

Once the software and workflows are preserved within the collaboration they are ready to be archived within the CERN Analysis Preservation Portal[13] (CAP), a in-progress project for the long-term analysis archival led by the CERN Scientific Information Services division. The portal is designed to hold a wide range of data regarding an analysis, including metadata such as a list of contributors, static data assets in the form of e.g. background estimates (that can be re-used for reinterpretation) as well as source code and software environments. The portal, based on the Invenio platform, will provide a web user interface to search and view preserved analyses.

3. Distributed Workflow Execution using modern Container Orchestration Engines

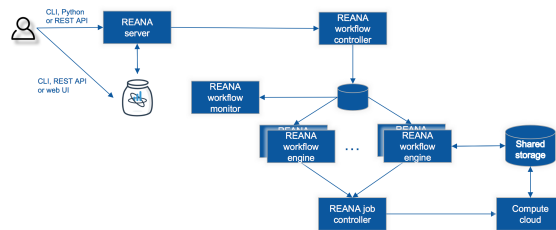
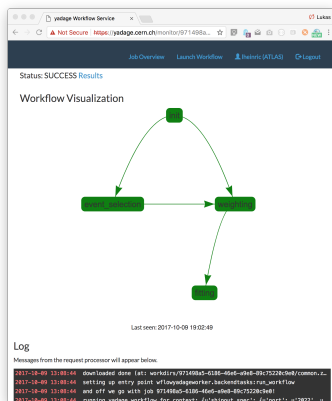


Figure 1. REANA workflow monitor and REANA architecture

With analysis software captured as standard OCI containers the workflows archived in CAP can be executed on distributed computing clusters with a variety of container orchestration tools, such as Kubernetes. These can be conveniently set up both on commercial cloud providers or using private clouds managed by OpenStack such as the CERN OpenStack Installation. In order to facilitate the execution of the workflows stored in the CERN Analysis Preservation Portal, a dedicated workflows-as-a-service platform REANA is being developed as a collaboration between CERN, DASPOS[14] and DIANA-HEP[15], offering collaboration members an easy entry-point to re-execute preserved analysis without managing a container cluster on their own. The platform will handle e.g. access control, storage (e.g. CephFS, EOS), CVMFS access (if needed), monitoring and logging based on industry and community standards. An overview of the architecture is shown in Figure 1.

```

type: recast_scan
title: 'Validation of ATLAS-CONF-2013-024 in Neutralino/Stop Mass Plane'
pubkey: 'cds/1525889'
request_format: 'standard_format'
description: >
  this reinterpretation re-validates the original grid in of ATLAS-CONF-2013-024
reason: >
  The grid is useful to validate / check third-party
  (such as CheckMate) implementations of the ATLAS-CONF-2013-024 as it
additional_information: >
  the grid consists of 36 points in the stop / neutralino mass plane.
  The input parameters are given in the standard format consisting of
  model parameters and number of events.parameters: [mStop, mNeutralino]
parameters: [mStop, mNeutralino]
points:
- coordinates: [200.0, 0.0]
  data: data/200_0_0.0.zip
- coordinates: [300.0, 0.0]
  data: data/300_0_0.0.zip
- coordinates: [300.0, 100.0]
  data: data/300_0_100.0.zip
...

```

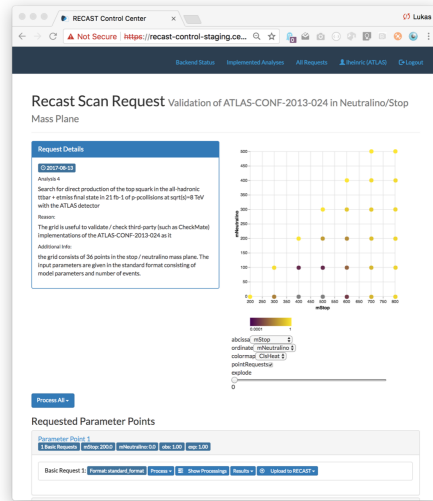


Figure 2. RECAST request and resulting web-based presentation

4. RECAST — leveraging preserved workflows for reinterpretation

Based on CAP as an analysis archive and REANA as an analysis execution engine, the collaboration-internal reinterpretation service can feasibly be implemented. This additional layer connects physics research questions to the generic but physics-unaware backends provided by CAP and REANA through a reinterpretation-specific interface and tools. Reinterpretations are well-posed scientific questions about BSM models that have statistical inference results as their answers. That basic structure is largely invariant to the specific implementing workflows for signal generation and analysis.

Therefore, a web-based interface is being developed to organize both models and inference results of reinterpretations. Reinterpretation requests can be uploaded via both the web interface as well via a command line tool that specifies the details of the reinterpretation such as the analysis to be reinterpreted and input data in the form of a file archive that holds parameter-point specific information. The specific format of the model input can be chosen between one of several options such as dataset identifiers for pre-generated signals available from collaboration-internal storage or Monte Carlo generation input data such as parameter cards and UFO model descriptions[16].

Depending on the input provided as well as the targeted analysis, the service then selects suitable implementations through a single workflow stored in CAP or a combination of workflows. The former is suitable if the input format is a pre-generated signal such that the analysis can be run directly on the new input, while the latter is chosen for the case where the analysis inputs need to be generated on the fly. In this case the composability of the workflow engine enables combining independently developed workflows for signal generation steps (different Monte Carlo generators, variable simulation settings) with analysis implementations that may include fully preserved ATLAS analyses but also re-implementations such as Rivet or collaboration internal truth-based analyses. The two main modes are visualized in Figure 3.

The desired combination can then be submitted to REANA for processing, where the workflow definitions are retrieved from a workflow repository like CAP and upon completion

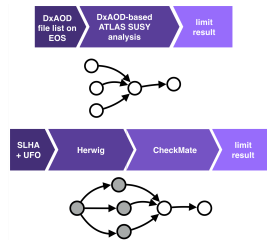


Figure 3. Workflows with and without signal generation.

the reinterpretation results such as limits or signal region event yields are transferred back to RECAST. Similarly to the model inputs, results may be produced by the workflows in a variety of ways (text files, JSON documents, ROOT files); RECAST converts those outputs into a standardized format, and presents them on the web interface and the associated command line tools.

5. Future Developments

Containerized workflows have been used for a number of ATLAS reinterpretations[17, 18] and based on these, the described system is being designed for streamlining the internal reinterpretation efforts of the ATLAS experiment by interfacing with the CERN Analysis Preservation portal. As the collaboration gains more experience with both analysis preservation and reinterpretation, it may be possible to open such a service to the wider high-energy community as envisioned by the RECAST proposal [19, 20]. In such a scheme phenomenologists could then upload reinterpretation requests. Depending on resource requirements and availability as well as scientific merit, the collaboration may then choose to fulfill such requests by reusing its preserved analysis and publishing the results on archives such as HepData[21].

References

- [1] Aad G *et al.* (ATLAS Collaboration) 2012 *Phys. Lett.* **B716** 1–29 (*Preprint* 1207.7214)
- [2] Aad G *et al.* (ATLAS Collaboration) 2008 *Journal of Instrumentation* **3** S08003 URL <http://stacks.iop.org/1748-0221/3/i=08/a=S08003>
- [3] Chatrchyan S *et al.* (CMS) 2012 *Phys. Lett.* **B716** 30–61 (*Preprint* 1207.7235)
- [4] Alves D (LHC New Physics Working Group) 2012 *J. Phys.* **G39** 105005 (*Preprint* 1105.2838)
- [5] Agostinelli S *et al.* 2003 *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** 250 – 303 ISSN 0168-9002 URL <http://www.sciencedirect.com/science/article/pii/S0168900203013688>
- [6] de Favereau J, Delaere C, Demin P, Giammanco A, Lematre V, Mertens A and Selvaggi M (DELPHES 3) 2014 *JHEP* **02** 057 (*Preprint* 1307.6346)
- [7] Buckley A, Butterworth J, Lonblad L, Grellscheid D, Hoeth H, Monk J, Schulz H and Siebert F 2013 *Comput. Phys. Commun.* **184** 2803–2819 (*Preprint* 1003.0694)
- [8] Drees M, Dreiner H, Schmeier D, Tattersall J and Kim J S 2015 *Comput. Phys. Commun.* **187** 227–265 (*Preprint* 1312.2591)
- [9] OCI 2017 Open Container Initiative URL <https://www.opencontainers.org>
- [10] Blomer J, Buncic P and Fuhrmann T 2011 *Proceedings of the First International Workshop on Network-aware Data Management NDM '11* (New York, NY, USA: ACM) pp 49–56 ISBN 978-1-4503-1132-8 URL <http://doi.acm.org/10.1145/2110217.2110225>
- [11] Kuncar J, Nielsen L H and Simko T 2014 Invenio v2.0: A Pythonic Framework for Large-Scale Digital Libraries Tech. Rep. ATLAS-CONF-2016-033 CERN Geneva URL <http://urn.fi/URN:NBN:fi-fe2014070432294>
- [12] Cranmer K and Heinrich L 2017 (*Preprint* 1706.01878)
- [13] Chen X, Dallmeier-Tiessen S, Dani A, Dasler R, Fernández J D, Fokianos P, Herterich P and Šimko T 2016 *CERN Analysis Preservation: A Novel Digital Library Service to Enable Reusable and Reproducible Research* (Cham: Springer International Publishing) pp 347–356 ISBN 978-3-319-43997-6
- [14] DASPOS 2017 DASPOS Project Website URL <https://daspos.crc.nd.edu/>
- [15] DIANA-HEP 2017 DIANA-HEP Project Website URL <http://diana-hep.org/>
- [16] Degrande C, Duhr C, Fuks B, Grellscheid D, Mattelaer O and Reiter T 2012 *Comput. Phys. Commun.* **183** 1201–1214 (*Preprint* 1108.2040)
- [17] Aad G *et al.* (ATLAS Collaboration) 2015 *JHEP* **10** 134 (*Preprint* 1508.06608)
- [18] Aaboud M *et al.* (ATLAS Collaboration) 2016 *JHEP* **09** 175 (*Preprint* 1608.00872)
- [19] Cranmer K and Yavin I 2011 *Journal of High Energy Physics* **2011** 38 ISSN 1029-8479 URL [http://dx.doi.org/10.1007/JHEP04\(2011\)038](http://dx.doi.org/10.1007/JHEP04(2011)038)
- [20] 2015 ATLAS Data Access Policy Tech. Rep. ATL-CB-PUB-2015-001 CERN Geneva URL <https://cds.cern.ch/record/2002139>
- [21] Maguire E, Heinrich L and Watt G 2017 *22nd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2016) San Francisco, CA, October 14-16, 2016* (*Preprint* 1704.05473) URL <https://inspirehep.net/record/1592380/files/arXiv:1704.05473.pdf>