

# Provenance tracking in the LHCb software

Ana Trisovic, Chris R. Jones, Ben Couturier, and Marco Clemencic

**Abstract**—Even though computational reproducibility is widely accepted as necessary for research validation and reuse, it is often not considered during the research process. This is because reproducibility tools are typically stand-alone and require additional training to be employed. In this paper, we present a solution to foster reproducibility, which is integrated within existing scientific software that is actively used in the LHCb collaboration. Our provenance tracking service captures metadata of a dataset, which is then saved inside the output data file on the disk. The captured information allows a complete understanding of how the file was produced and enables a user to reproduce the dataset, even when the original input code (that was used to initially produce the dataset) is altered or lost. This paper describes the implementation of the service and gives examples of its application.

**Index Terms**—data provenance, reproducibility



## 1 INTRODUCTION

REPRODUCIBILITY of research is accepted as necessary for result validation, sharing and reuse [1], [2]. Dozens of standalone tools have been created to preserve research materials and thus enable reproducibility. Some of these tools help with documentation [1], the others with tracking system dependencies thus allowing cross-platform compatibility [3], some even capture the whole analysis in an automatized workflow [4]. A number of reproducibility tools are web-based with a user-friendly graphical interface [5]. When employed in the research process, they are indeed able to solve the prob-

lems of reproducibility and allow reuse. However, reproducibility-focused research is still not common in scientific communities, and a small number of people who use the tools represent an exception. There are several reasons why these tools, although useful, are not ubiquitously used. First, researchers are often working under constant pressure to present results and publish papers [6], which can cause lack of time to employ reproducibility practices in their everyday work. Often in research groups, many computationally intensive tasks are done by students or young scientists with typically fixed-term stay, during which they might not manage to ensure reproducibility of their work. Second, there is often a lack of training on best practices in data analysis in social and (some) natural sciences, even though these disciplines also face an expansion of data and the need for computationally intensive analyses. Researchers in those fields are hence hesitant to use more software than they need to, including the reproducibility tools. Because of this, the tools are not normally used in everyday scientific work, which can thus hinder research

- 
- *A. Trisovic was with the University of Chicago, 1100 E 57th St, Chicago, IL 60637 and CERN - European Organization for Nuclear Research 1, Esplanade des Particules, CH-1217 Meyrin, Switzerland.  
Email: anatrissovic@g.harvard.edu*
  - *C. R. Jones is with the Cavendish Laboratory, University of Cambridge, JJ Thomson Ave, Cambridge CB3 0HE, United Kingdom.*
  - *B. Couturier and M. Clemencic are with CERN - European Organization for Nuclear Research 1, Esplanade des Particules, CH-1217 Meyrin, Switzerland.*

reproducibility and reuse at a later stage.

In this paper, we argue that the best way to foster reproducibility is to integrate it within existing scientific software that is already in use. This way enabling reproducibility is simple and it seamlessly integrates into daily researchers' work. It lessens the burden of the analysts who might otherwise need to install and learn to use a new software tool. We demonstrate the proposed solution through our work at LHCb, the high-energy physics experiment at CERN. According to a survey conducted at the experiment [7], most LHCb analysts use the official software to fully or in part perform their physics analysis. These results served as a starting point for addressing the reproducibility challenge. Our solution is designed and developed within the official experimental software to capture data provenance, which is then saved inside the output data file on the disk. The stored provenance allows understanding how a file was produced [8] and provides sufficient information to entirely reproduce the dataset, eliminating the need for the original input code or even documentation.

## 2 THE LHCb SOFTWARE

The LHCb software is the essential component in all data-related activities at LHCb. It is used in a range of data processing environments, from real-time data collection in the experimental setup, data reconstruction, simulation, to advanced physics analyses and collision visualization. The LHCb software is based on the C++ object-oriented framework called GAUDI [9], which provides a common infrastructure and environment for the software applications of the experiment. GAUDI was used during both the first and second run of the LHC and it proved to be a reliable and flexible platform for LHCb data processing on the CERN Computing Grid. The project was created in 1998 by the LHCb collaboration, but since then it was released open-source<sup>1</sup> and adopted by a number of other high-energy physics experiments such as ATLAS [10].

1. <https://gitlab.cern.ch/lhcb/Gaudi>

The software is organized in a modular architecture of smaller and more manageable packages. Each module has a defined functionality and an interface through which it interacts with the other components. Such architecture provides layers of abstraction for developers, meaning that one does not need to understand the whole framework to contribute to one of the components.

The basic modules of the GAUDI framework and the connections between them are shown in Figure 1. The Application manager (in code denoted as `ApplicationMgr`) controls the execution of the jobs within the framework. It creates and initializes the required modules in the system, and retrieves input data. The input data is a collection of highly-structured information that describe particle collisions (also called events) recorded inside the detectors or created in simulations. The Application manager loops over the input data events and executes the algorithms. If there are any errors in the execution, the Application manager will handle them and in the end it will terminate the job. The module Algorithms (`Algorithms`), has a central place in the job execution, as it performs data processing on all input data. At the end, output files are produced in a form of another data file or other type of output.

The GAUDI services provide various utilities and services for the Algorithms in the system, which are also initialized by the Application manager at the beginning of a job. Normally, only one instance of a service is required in the job. There are a number of different services within the framework that can be used by the Algorithms but some of the main ones are:

- Event Data service (`EventDataSvc`) and Histogram service (`HistogramDataSvc`) that read and process individual collision events,
- Detector Data service (`DetDataSvc`) for capturing detector data,
- Message service (`MessageSvc`) logs progress or errors in the Algorithms and
- Tool Service (`ToolSvc`) manages Algorithm tools, which are required during the

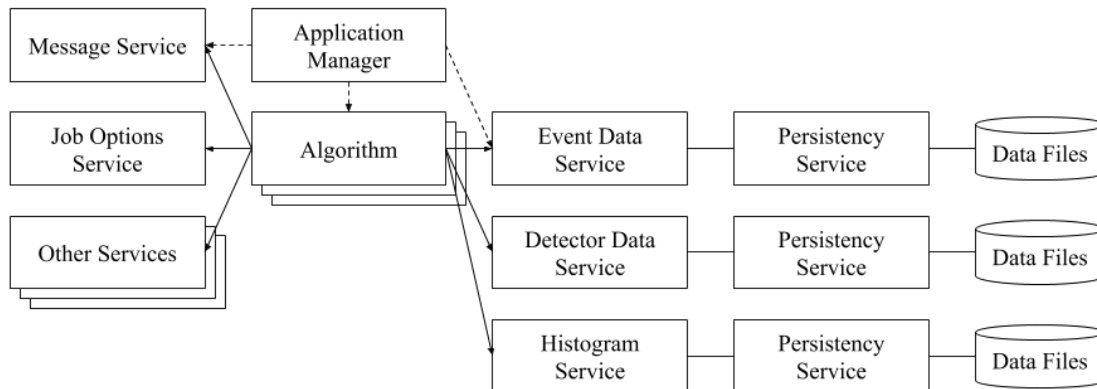


Fig. 1. A snapshot of GAUDI during runtime shows the basic modules and the connections between them.

### Algorithm execution.

The Persistency services allow writing the output data on the disk as presented in the figure. Finally, there are many other services in the framework that provide specialized functions that can be enabled and disabled by the users. Each of the service classes is used by the Algorithms via an interface, which is a helper class that defines the functionality of a service through a number of public methods. These methods also allow the communication of the service with other components of the framework.

In order to initiate these services, GAUDI ingests a set of configurations defined by physics analysts in (one or more) Python files at the beginning of a data processing job. The Python files are called *Python application configuration files* and they specify how the job will run and designate configurations to required C++ objects in GAUDI. These configurations are passed by the Job Options Service (`JobOptionsSvc`) and applied at runtime. Every Algorithm used in the framework is configured using these options.

### 3 IMPLEMENTATION OF THE PROVENANCE TRACKING SERVICE

The provenance tracking service, called Metadata service (`MetaDataService`), is implemented as a GAUDI service inside the module GAUDI services

(`GaudiSvc`). Its functionality is simple: collect information about a job and capture it in an object, which is then stored as metadata in the output data file. There is a number of LHCb data formats in which an output of a GAUDI job can be stored. Data formats that are most commonly used in high-energy physics analysis are based on the ROOT file format [11]. A ROOT file format is very flexible and it acts like a UNIX file directory, meaning that it can store directories and data objects organized in an arbitrary order. Furthermore, it can store any C++ object, like for example histograms, plots, tables and other. The metadata object, named *info*, is implemented as a dictionary (a `map std::map` in C++), which is a general-purpose data structure for storing a group of objects. A dictionary captures a set of key-value pairs, which, in this implementation, are names of application properties as keys, and their configurations as values.

The service is implemented in a C++ class with the following main methods that execute its workflow in runtime:

- 1) `isEnabled` captures information whether the service is enabled in the job or not.
- 2) `start` initiates the service and calls `collectData`.
- 3) `collectData` executes the main

functionality of the service. It traverses and queries the GAUDI tools (`ToolSvc`), services (`Services`), algorithms (`Algorithms`) and Job Options (`JobOptionsSvc`) to capture their configurations.

- 4) `getMetaData` returns the object that stores a dictionary of the job configurations.

Within the GAUDI framework, there are three audit methods that follow a job execution. They are automatically invoked by the Application manager at the start of every job. Those methods are:

- `initialize` that initializes algorithms and services, and applies job options,
- `execute` that executes the main function of the job,
- `finalize` that is called at the end of the job.

The Metadata service is called and initialized from the `finalize` method, at the point when all configurations have been correctly applied to the job. During the execution of the `initialize` or `execute` methods only a selection of the job configurations (not all) have been applied, meaning that calling the service from these methods would cause an information loss. Therefore, the metadata is only captured at the end of the job execution at the moment when the output ROOT file is written to the disk. The service functionality is finished when the metadata dictionary is saved into the file.

## 4 USING THE PROVENANCE TRACKING SERVICE

Even though LHCb data analyses can be done using a wide range of tools and programming languages, retrieving the data from the CERN Computing Grid needs to be done using the LHCb software. Typically an LHCb application called DAVINCI is used in this step. DAVINCI is a physics analysis application based on the GAUDI framework that is primarily used as a part of data processing to calculate a variety of

kinematic quantities for recorded particles, but it is also used in fine data selections to extract particle decays of interest. As any other GAUDI-based physics application that is used in LHCb, DAVINCI can access and utilize the Metadata service available in the framework.

Capturing data provenance within the GAUDI framework is the simplest way for the analysts to conduct reproducible analysis. The provenance captured with the Metadata service is useful in a number of different scenarios:

- 1) The first scenario refers to reproducing a dataset using the original application version. A common practice in using the LHCb software is to use the latest available version at the time, which is done by specifying the keyword 'latest' in the code. The latest version is recommended for use because it captures recent developments with new features or solutions to known bugs. However, by doing this the analysts do not necessarily record what application version they used, which may at the later stage, when the "latest" version changes, hinder reproducibility. Even though the applications run on the same framework, two different versions can produce different datasets. For example, the application DAVINCI v39R0 and DAVINCI v42R3 may produce slightly different outputs even when they ingest the same application configuration file. The Metadata service hence facilitates reproducibility by storing the exact information about the version of the software used, which users can unequivocally specify when reproducing the dataset.
- 2) Another common practice in conducting physics analyses is to use one application configuration file to produce many similar datasets with slight configuration changes. This is typical when for example an analyst wants to test different selections on the data. Such practice produces a number of data files without clear information how they were produced, and if

- an analyst by mistake mislabels them (by marking, for example, the origin year of data 2012 instead of 2011) mistakes in the physics analysis may happen. The Metadata service facilitates reproducibility in this scenario as it provides information (that is not prone to human error) about these slight configuration changes in the dataset and thus it enables the analyst to double-check their work.
- 3) Furthermore, working with collaborators on an analysis typically means sharing disk workspace on a CERN or University server. Each of the collaborators creates temporary or derived data files, and often after some time they forget how and why these files were created. This hinders reproducibility as it conceals potential steps that were previously taken in the analysis workflow. The Metadata service captures what were the inputs and processes conducted on each dataset, which could help the analysts to reconstruct the analysis workflow and distinguish between temporary (testing) and requisite (final) datasets.
  - 4) Finally, if a bug in one application version is identified, the analysts need to know whether they had created datasets using this version as it could negatively affect their analysis results. If the provenance of these datasets is available, they could instantly evaluate whether they used the faulty version and recreate the datasets with another application version. In this scenario, the information provided by the Metadata service helps in troubleshooting, result verification and ensuring trustworthiness.

The Metadata dictionary provides clear information what application version was used and how it was used, thus avoiding the ambiguity illustrated in these scenarios. Our solution can be immensely useful for everyday development and validation of research.

In the current implementation, the service needs to be manually enabled in order to be used

in data processing. This is done by including just one additional line of code in the Python application configuration file, that appends the Metadata service into the list of existing services handled by the Application manager. Once the Python application configuration file is completed, it is passed to GAUDI to run the job.

An output dataset in the ROOT file format that captures data of a particle decay  $D^+ \rightarrow K^+\mu + \mu^-$  (D2Kmumu) is shown in Figure 2. The dataset was created using the application DAVINCI V42R3. This information and other configurations are captured in the additional *info* object, which is also saved and visible in the figure.

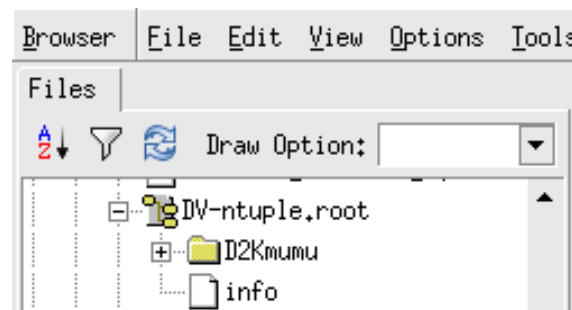


Fig. 2. After data file DV-NTUPLE.ROOT has been produced with the provenance tracking service, it stores an additional object with the job metadata called *info*, which can be seen in this file tree.

The *info* file can be read in two different ways. The first one is through command-line, by simply reading in and printing the content of the dictionary. The second way is to view the content of the *info* file from a stand-alone provenance viewer, as shown in Figure 4. The viewer is implemented as a pop-up window based on C++ and the ROOT programming language. This means that it requires the ROOT application (and an input ROOT file) to be executed. The viewer uses a table to showcase the dictionary (key-value pairs) of metadata captured in the job. It allows for a user-friendly way to see the provenance, while the command-line approach is better when there is a need to reuse the configurations or reproduce the dataset.

The screenshot shows a window titled "Meta Data Viewer" with a table of settings. The table has two columns: the first column lists application manager attributes, and the second column shows their corresponding values. The values are displayed in blue text within input fields.

Attribute	Value
ApplicationMgr.ActivateHistory	False
ApplicationMgr.AlgTypeAliases	{ }
ApplicationMgr.AppName	DaVinci
ApplicationMgr.AppVersion	v42r3
ApplicationMgr.AuditAlgorithms	False
ApplicationMgr.AuditServices	False
ApplicationMgr.AuditTools	False

Fig. 3. The HTML-based graphical user interface of the Metadata viewer.

When it is necessary to reproduce a ROOT dataset, the original job can be recreated from the information within the dataset. This is done by extracting the metadata from the dataset and saving it in a file as a “flat” list of options. The list is essentially a sequence of job configurations and it will resemble the original Python application configuration file(s) created by the physics analysts. The list of configurations can be saved as a Python file, Python pickle file or Linux configuration options file. These file formats are typically used for serializing Python objects. GAUDI understands and processes such a file (the flat list) in the same way as the original application configuration file. Furthermore, the original application configuration file is no longer needed. Even though DAVINCI is most commonly used for physics analyses at LHCb, the Metadata service can be used for other applications within the framework that can process and produce ROOT datasets. The service was first released in GAUDI v27R1, meaning that it is available for use in physics analyses with DAVINCI v40R0 onward.

## 5 CONCLUSION

In this paper, we introduced a new functionality of the LHCb software framework GAUDI that captures provenance of a job and stores it directly within the output dataset. Using only

the stored provenance of a dataset, the original job and the dataset can be independently reproduced. We argue that the most effective way to integrate reproducibility is through the solutions that easily incorporate into researchers’ daily workflows. Our solution is implemented within an existing and widely-used scientific software. Because it does not require the analysts to conduct additional work like install third-party software or learn a new skill, it lightens the effort of enabling reproducibility in physics analyses at LHCb. In the second part of the paper, we presented our implementation and explored several different use-cases in which the provenance service would facilitate research documentation and reproducibility.

## ACKNOWLEDGMENTS

Ana Trisovic acknowledges funding from the CERN Doctoral Student program. During her time at the University of Chicago, she was funded in part by the Sloan Foundation.

The authors would like to thank Freddie Witherden and Patrick O’Leary, who conducted peer-review of this paper, for their insightful feedback and comments. Their comments are public and can be found at [10.22541/au.158014063.32681377](https://arxiv.org/abs/10.22541/au.158014063.32681377) and [10.6084/m9.figshare.11536461](https://arxiv.org/abs/10.6084/m9.figshare.11536461).

## REFERENCES

- [1] X. Chen, S. Dallmeier-Tiessen, R. Dasler, S. Feger, P. Fokianos, J. B. Gonzalez, H. Hirvonsalo, D. Kousidis, A. Lavasa, S. Mele *et al.*, “Open is not enough,” *Nature Physics*, p. 1, 2018.
- [2] E. National Academies of Sciences, Medicine *et al.*, *Reproducibility and replicability in science*. National Academies Press, 2019.
- [3] C. Boettiger, “An introduction to docker for reproducible research,” *ACM SIGOPS Operating Systems Review*, vol. 49, no. 1, pp. 71–79, 2015.
- [4] J. Goecks, A. Nekrutenko, and J. Taylor, “Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences,” *Genome biology*, vol. 11, no. 8, p. R86, 2010.
- [5] T. Staubitz, H. Klement, R. Teusner, J. Renz, and C. Meinel, “Codeocean—a versatile platform for practical programming exercises in online environments,” in *Global Engineering Education Conference (EDUCON), 2016 IEEE*. IEEE, 2016, pp. 314–323.
- [6] M. Baker, “Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the crisis rocking science and what they think will help,” *Nature*, vol. 533, no. 7604, pp. 452–455, 2016.
- [7] A. Trisovic, “Data preservation and reproducibility at the lhcb experiment at cern,” Ph.D. dissertation, University of Cambridge, 2018.
- [8] T. Pasquier, M. K. Lau, A. Trisovic, E. R. Boose, B. Couturier, M. Crosas, A. M. Ellison, V. Gibson, C. R. Jones, and M. Seltzer, “If these data could talk,” *Scientific data*, vol. 4, 2017.
- [9] G. Barrand, I. Belyaev, P. Binko, M. Cattaneo, R. Chytrcek, G. Corti, M. Frank, G. Gracia, J. Harvey, E. Van Herwijnen *et al.*, “GAUDI - A software architecture and framework for building HEP data processing applications,” *Computer Physics Communications*, vol. 140, no. 1, pp. 45–55, 2001.
- [10] M. Clemencic, H. Degaudenzi, P. Mato, S. Binet, W. Lavrijsen, C. Leggett, and I. Belyaev, “Recent developments in the LHCb software framework Gaudi,” *Journal of Physics: Conference Series*, vol. 219, no. 4, p. 042006, 2010. [Online]. Available: <http://stacks.iop.org/1742-6596/219/i=4/a=042006>
- [11] R. Brun and F. Rademakers, “ROOT—an object oriented data analysis framework,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 389, no. 1-2, pp. 81–86, 1997.

## ABOUT THE AUTHORS

**Ana Trisovic** is a Sloan postdoctoral fellow at the Institute for Quantitative Social Science, Harvard University. Her main research interests are on computational reproducibility, data provenance and open science. She completed her PhD

in computer science at the University of Cambridge on the topic of data preservation and reproducibility. During that time, she worked on the LHCb experiment, CERN Open Data and CERN Analysis Preservation. Contact her at [anatrivic@g.harvard.edu](mailto:anatrivic@g.harvard.edu) and find her website at <https://anatrivic.com>.

**Chris R. Jones** is a Senior Research Associate at the Cavendish Laboratory at the University of Cambridge. He completed his PhD in particle physics at Magdalene College at the University of Cambridge. He oversees the work and development of the RICH detector at the LHCb experiment. Contact him at [jonesc@hep.phy.cam.ac.uk](mailto:jonesc@hep.phy.cam.ac.uk)

**Ben Couturier** is a software engineer at CERN. He develops software for the LHCb experiment. He is part of the team putting in place tools and processes to ensure software quality, with a special focus on preservation. Contact him at [Ben.Couturier@cern.ch](mailto:Ben.Couturier@cern.ch)

**Marco Clemencic** is a lead software developer for the Gaudi framework at the LHCb experiment. He completed his PhD in particle physics at the University of Turin. Contact him at [Marco.Clemencic@cern.ch](mailto:Marco.Clemencic@cern.ch)