Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Open Science with the LHCb Ntupling Service

Christine A. Aidala
University of Michigan
Michael Buchar
European Organization for Nuclear Research
Chris Burr
European Organization for Nuclear Research
Dillon S. Fitzgerald
University of Michigan
Cameron Duncan McClymont
European Organization for Nuclear Research
Adam Morris
Vilnius University
Sebastian Neubert
University of Bonn
Piet Nogga
pnogga@uni-bonn.de

University of Bonn
Ryunosuke O'Neil
European Organization for Nuclear Research
Eduardo Rodrigues
University of Liverpool
Daan Eduard Rosendal
Now at Lakera company
Tibor Simko
European Organization for Nuclear Research

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11445690/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)



October 20, 2026

Open Science with the LHCb Ntupling Service

Christine A. Aidala[1], Michael Buchar[2], Chris Burr[2], Dillon S. Fitzgerald[1], Cameron Duncan McClymont[2], Adam Morris[3], Sebastian Neubert[4], Piet Nogga[4], Ryunosuke O'Neil[2], Eduardo Rodrigues[5], Daan Eduard Rosendal[6], Tibor Šimko[2]

[1] *University of Michigan, United States*
[2] *CERN, Switzerland*
[3] *Vilnius University, Lithuania*
[4] *University of Bonn, Germany*
[5] *Oliver Lodge Laboratory, University of Liverpool, Liverpool, United Kingdom*
[6] *Now at Lakera company, Zurich, Switzerland*

Abstract

The data samples collected at the Large Hadron Collider at CERN are highly valuable, and the analysis of these data has led to numerous discoveries and the 2025 Breakthrough Prize in Fundamental Physics. Preserving and providing open access to these data for future analysts is imperative to maximize the scientific output of the collected data. The LHCb collaboration has developed a novel approach to this pursuit of open science, the LHCb Ntupling Service: a web service for on-demand production and publishing of custom LHCb open data, providing open access to both Run 1, and for the first time, Run 2 proton-proton collision data collected by the LHCb experiment, amounting to over 4 PB of data to explore.

Submitted to EPJ Research Infrastructures

© 2026 CERN for the benefit of the LHCb collaboration. CC BY 4.0 licence.

1 Introduction

Contemporary research in high energy physics (HEP) necessitates vast data volumes stored in custom file formats to meet the needs of experimental programs. This typically involves experiment-specific software to parse the custom file formats and extract necessary information for physics analysis. Both the complexity and sheer volumes of these datasets make it challenging to preserve and provide open access to these data. For example, all four main Large Hadron Collider (LHC) experiments have differing computing models and resulting file formats [1–8], underlining the intricacies of the data preservation efforts in HEP. Nevertheless, data preservation efforts are crucial to maximize the physics potential of the collected samples. For this, CERN is committed to providing open access to these data as outlined in the CERN Open Data Policy [9], encouraging adoption of a consistent approach between the four LHC experiments to publish scientific data in line with standards for Findability, Accessibility, Interoperability, and Reusability (FAIR) [10].

At the end of 2022, LHCb announced the LHCb Ntuple Wizard concept [11], a web application for configuring ntupling jobs for the LHCb Analysis Productions batch processing system [12] to be run on the Worldwide LHC Computing Grid (WLCG). In this case, ntupling refers to processing and storing experimental data into a columnar data format suitable for analysis with standard tools without the need for experiment-specific software. Significant progress has led to the development of an overarching system, the LHCb Ntupling Service [13, 14], to seamlessly handle the interface between the LHCb Ntuple Wizard and Analysis Productions, with dedicated databases for details of user requests and resulting ntuples from production jobs. It is an all-in-one web service for users to request custom ntuple productions, track the status of the ntupling jobs, communicate with the open data team, and retrieve the resulting ntuples. Internally, ntuple production at LHCb is handled by the `DaVinci` application [1,2,15] through Python scripts with specific syntax related to the LHCb data format. A noteworthy advantage of the LHCb Ntupling Service is that configurations needed to steer the `DaVinci` application are captured through the Ntuple Wizard web interface, removing the necessity of learning `DaVinci` to produce ntuples. Throughout the paper, the LHCb Ntupling Service will often be referred to as the "service", while the LHCb Ntuple Wizard will often be referred to as the "wizard".

1.1 Current Open Data Landscape

The four large LHC experiments, in addition to many other HEP experiments, have efforts to provide open access to their collected data, each with their own computing model, leading to a complex landscape of data formats to support. Open data from HEP experiments are typically provided to users in one of two ways:

1. Experiment-specific data formats, along with the necessary software and instructions to analyze these data.

2. A derived data format, such as a ROOT [16, 17] ntuple[1], removing the need for experiment-specific software.

[1]For each physics object or event of interest, a tuple of information is saved with n columns corresponding to n different physical quantities.

Both approaches have their advantages. Approach (1) offers additional flexibility and customization at the cost of a significantly steeper learning curve, requiring custom software and tools, as well as sufficient CPU to parse the data. Approach (2) is typically limited to particular selections meant to isolate physics processes of interest, ultimately up to the collaboration releasing the data. While this is excellent in terms of reducing the barrier of entry to work with these data, the fine-grained choices of physics objects and selections provided by collaborations can stifle creativity in research using open data. On the other hand, the approach in (1) comes at a significantly higher resource cost, as full data sets stored in experiment-specific file formats typically have large data volumes that need to be duplicated for public access. In December 2023, LHCb released the entire Run 1 dataset to the CERN Open Data Portal (ODP) [18], amounting to about 800 TB of collision data in custom file formats, corresponding to approach (1). Applying the same methodology to the Run 2 data would amount to roughly 10 PB of custom file formats copied to the CERN Open Data Portal, which was not feasible from the standpoint of resource availability.

It is worth mentioning the significant efforts by other HEP experiments to generalize the approach in (2) to provide a derived data format rich in high-level event information suitable for a number of physics analyses, without the need for experiment-specific software. Both the CMS and ATLAS collaborations at the LHC have efforts in this direction with their so-called NanoAOD [19] and PHYSLITE [20] data formats, respectively, both with accompanying open data releases [21, 22]. However, reduction of average event size is typically accomplished by isolating particular physics objects and/or skimming away event-level information.

The novel web service presented in this paper introduces an innovative methodology for providing open access to data collected in HEP experiments. The LHCb Ntupling Service allows for on-demand production of custom LHCb open data, offering exploration of the LHCb data streams, specification of physics object selection, and customization of information to be written in the resulting ntuple(s). This offers both flexibility in terms of exploring physics selections, and a lower learning curve without the need for experiment-specific software to process the resulting ntuples, capturing the best of approaches (1) and (2). It also results in significantly smaller data volumes on the CERN Open Data Portal, as produced ntuples are typically on the order of a few to hundreds of GB.

The paper is organized as follows: the architecture of the LHCb Ntupling Service is discussed in Section 2, followed by a detailed look at the user experience in Section 3. The integration with the CERN Open Data Portal is presented in Section 4, followed by a discussion of future plans / outlook in Section 5, and finally, a summary in Section 6.

2 Architecture

The LHCb Ntupling Service comprises three code repositories: a React [23] **frontend** user interface, a Representational State Transfer (REST) **backend**, and a **DevOps** repository holding the deployment tooling and Helm chart, all privately hosted on the open GitLab code repository platform [24]. A fourth component, the **LHCb Ntuple Wizard** [11], is consumed by the frontend and is core to the building and submission of ntupling requests (see Section 2.2). The LHCb Analysis Productions batch processing system [12] is the engine behind running production jobs on the WLCG to fulfill the requests, which

are themselves stored as GitLab issues in a private repository comprising the requests database. A diagram of the system architecture is shown in Figure 1.

2.1 Component Topology

The service has a three-layer structure:

- Incoming traffic first reaches a CERN Single Sign-On (SSO) proxy, which handles authentication and forwards requests to the frontend. Non-CERN users have options to sign in with a guest email, their organization (eduGain), or with a social account such as Google, GitHub, Facebook, Apple, or LinkedIn. CERN users can simply sign in using their CERN account.
- The frontend is a React web application that uses Nginx [25] to serve files and reverse-proxy `/api` requests to the backend.
- The backend is a Go [26] server that uses GitLab to store and track requests.

Each ntupling request is a GitLab issue in a private repository. The issue's labels encode the request status, the description includes details about the requester, and the comment thread facilitates discussion between the requester and the LHCb team. This design outsources the state management, audit trail, and human review workflow to GitLab's existing infrastructure rather than writing our own bespoke system.

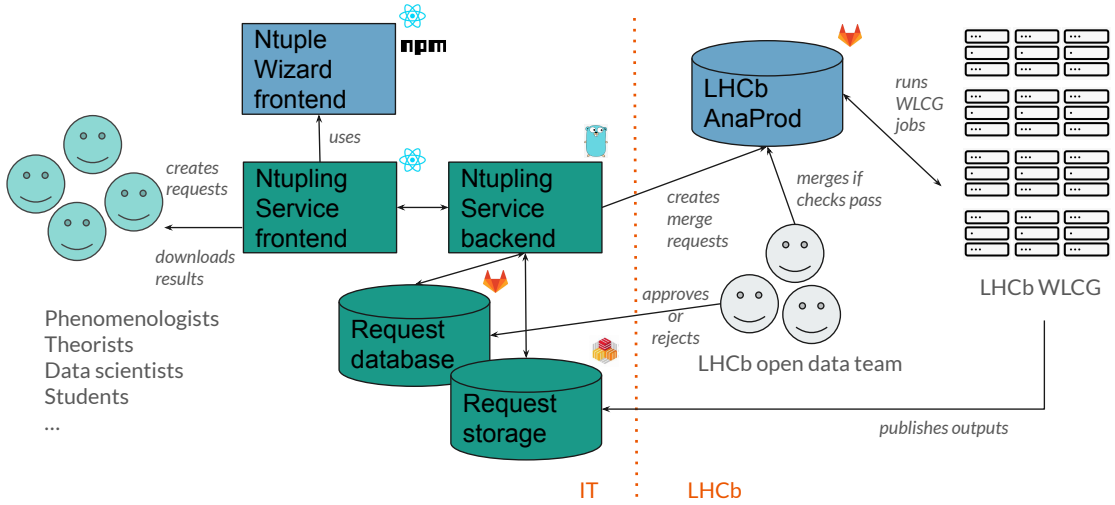


Figure 1: Architecture of the LHCb Ntupling Service, integrating the LHCb Ntuple Wizard and the Analysis Productions batch processing system. The rectangles represent code repositories while the cylinders represent databases, with light blue and green shading representing components managed by LHCb and CERN IT, respectively. The logos identify the technologies used for the different components, and the orange dashed line represents the operational division between LHCb and CERN IT. While LHCb is responsible for maintaining the wizard, CERN IT handles the integration with the service.

2.2 Ntuple Wizard Integration

The request creation flow is provided by the **LHCb Ntuple Wizard**, a separate component embedded into the frontend via its npm package (`lhcb-ntuple-wizard`) [27,28]. The wizard itself is hosted on GitLab [29], where it is tagged and versioned for corresponding releases on npm. This keeps the logic of assembling a valid ntuple configuration decoupled from the request management logic, and allows for independent development and reuse of the wizard.

The wizard guides the user through selecting datasets and physics objects of interest (e.g. decay candidates), and generates the necessary `YAML` files to steer the production configuration client-side. Upon submission, it sends the generated files to the backend, where they seed the request issue. After review from the LHCb Open Data team, the committed configuration is transferred to Analysis Productions to fulfill the request.

2.3 Storage

Two storage systems support the service:

- A **CephFS** file share [30] for user profiles, comments, and bookkeeping; and
- A set of **EOS directories** for the large ntuple output files and the gallery of curated example requests, where EOS is CERN's multi-PB disk-only storage technology [31].

2.4 Request Lifecycle

Each request flows through a set of statuses modeled as GitLab issue labels as shown in Figure 2.

A request may terminate for one of several reasons: failure during the running of the test/full production; cancellation by the requester after reviewing the test production results; or rejection by the LHCb team after review if a problem with the configuration is spotted. The LHCb team can approve/reject requests from GitLab by changing issue labels. Most of the other status changes are event-driven and handled automatically by the Ntupling Service using GitLab's API.

The test production runs over a small subset of the selected data stream(s), and the resulting ntuples are returned to the user upon completion so that the user can inspect them to verify that the desired variables are available, and use them to set up and test their analysis workflow. After reviewing the test production results, the user can either:

- Proceed to triggering the full production; or
- Make the appropriate modifications, and submit a new request.

For the latter case, a feature to clone configurations of past requests can be used as a starting point. For the former, the LHCb Open Data team is informed and performs a final inspection before merging the configuration files into the Analysis Productions repository to trigger the full production. Upon completion of the full production, the ntuples are copied to a location in EOS which is exposed to the service backend. The user can then view and download the result files from the web interface. Users are updated by email upon request status changes and when the LHCb team comments on their request(s).

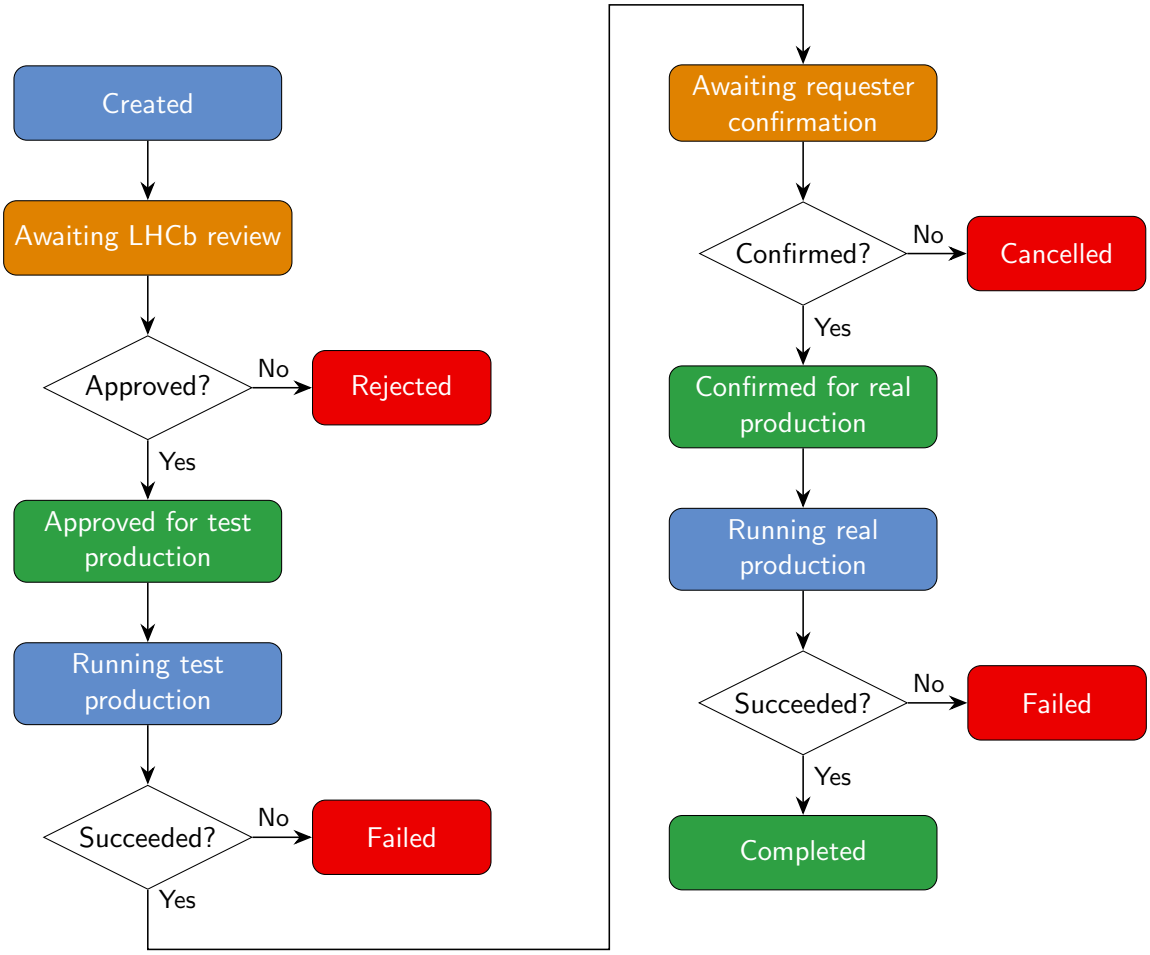


Figure 2: Flowchart of GitLab issue labels used to process user requests.

2.5 Security Measures

Security measures have been put in place to ensure that no malicious code can be injected into the WLCG or the LHCb file system through the service.

Configuration of production jobs from the wizard is captured in static `YAML` files. The service never accepts executable code from the user. Rather, the backend itself generates the necessary Python production files from a request's `YAML` files, so the code that runs in Analysis Productions originates from the Ntupling Service and not from the user. The backend also ensures that only the expected file types are received, and that configuration files have not been modified to access unreleased data samples. Because the backend also writes files to shared storage using user-supplied filenames, these are also validated to prevent directory-traversal attacks.

User authentication is delegated to CERN's SSO proxy, which injects the verified user email into the `X-Forwarded-User` header [2] read by the backend. Because the backend can only be reached by going through the SSO proxy, this header cannot be forged externally. In addition, operations on a request (commenting, canceling, approving results, etc.) are protected by a check comparing the SSO-asserted email against the requester's email in

[2] `The X-Forwarded-User` request header allows the identification of the originating IP address of a client connecting to a web server through a proxy server.

the GitLab issue, so users can only act on their own requests.

Finally, the LHCb team retains manual control over much of the request workflow. No request is approved automatically: the team reviews each one by hand and controls when to run it on Analysis Productions, rather than being triggered by the service itself. While the safeguards above defend against known attack surfaces related to malformed/malicious configurations, this human review provides an extra layer of security, ensuring that nothing runs on the WLCG without a person having inspected and approved it.

3 User Experience

A core design principle of the LHCb Ntupling Service is to provide an intuitive and guided user experience. In developing the web service, particular emphasis was placed on lowering the barrier to requesting LHCb open data as much as possible. Sections 3.1 and 3.2 describe the measures taken to support this goal, including comprehensive documentation and in-application guidance that help users understand the data they are requesting. Sections 3.3 and 3.4 discuss the primary communication channel for user support, as well as initial feedback and usage statistics for the LHCb Ntupling Service. Finally, Section 3.5 summarizes different use cases encountered for the service so far, and feedback obtained from users.

3.1 Guiding a First-Time User

Upon first visiting the website, users are met with a welcome message, and the opportunity to visit a growing gallery of example requests with corresponding ntuples for inspiration, or to create their own request, as shown in Figure 3. An expanded view of the $B^{\pm}$ meson mass gallery example is shown in Figure 4. Gallery examples are also accompanied by open data records on the CERN Open Data Portal, with Digital Object Identifiers (DOIs), citation templates and a detailed usage guide [32] on how to analyze the example data including REANA [33] reproducible analysis workflows.

The usage guide provides step-by-step instructions for using the web service, from request creation, the selection of LHCb datasets, physics objects and output variables, to submission and progression tracking. It also explains common LHCb nomenclature throughout the documentation. Additionally, the various pages encountered when configuring or viewing requests link directly to relevant sections of the guide. The usage guide also hosts analysis examples for the gallery requests, providing easy access to ROOT and Python code that demonstrates typical LHCb workflows, from reading and filtering ntuples to performing downstream analysis.

3.2 Creating a Custom LHCb Open Data Request

When preparing a new request, users first select the available physics objects, typically by choosing a physics decay process. At this stage, the search can already be constrained by selecting tags, particles contained in the decay, or so-called *Stripping* selection lines [34], which correspond to the output of the LHCb preselection processing. This interface is shown in Figure 5. As illustrated in Figure 6, the configuration of the requested ntuple(s) is fully customizable and relies on LHCb-internal tools to write the desired variables to the

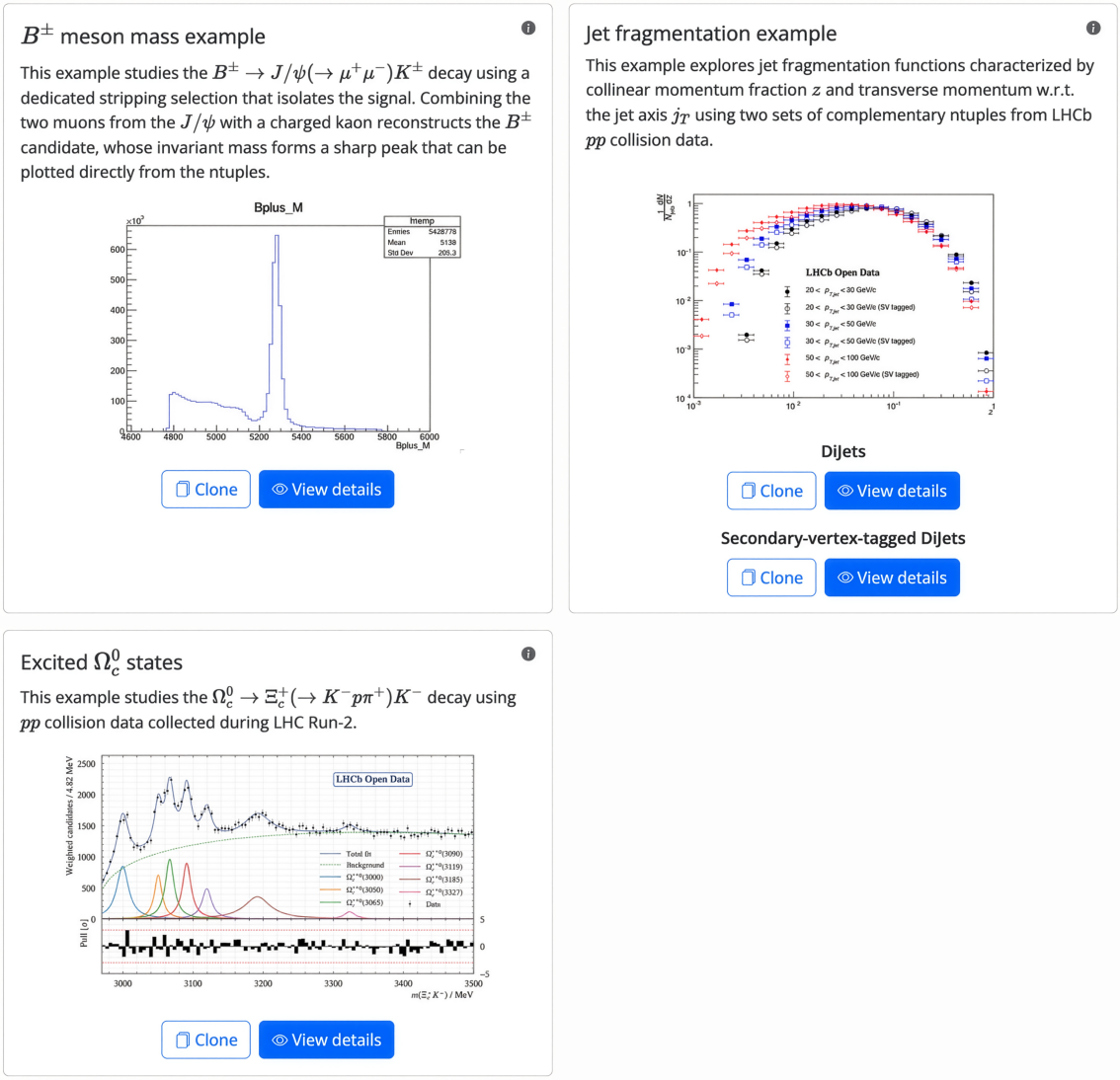


Figure 3: The gallery page for viewing example requests, working with preproduced ntuples, and / or cloning the corresponding request configuration as a template for a new request.

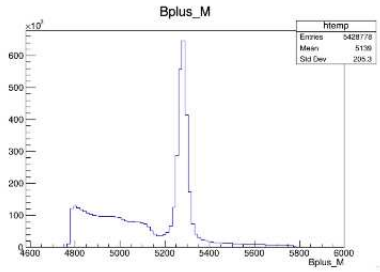
output ntuple(s). A set of default tools is provided to include commonly used variables, such as information about particle kinematics and geometry, and operational metadata from the LHC machine and the LHCb experiment. Finally, users specify the data to be processed by LHCb on the WLCG on their behalf. The available datasets are categorized into physically meaningful subsets of the collision data, referred to as *Streams*, as well as by data-taking year and LHCb magnet polarity. This step is shown in Figure 7. Detailed information regarding the request generation procedure can be found in Ref. [11].

3.3 Communication and Request Tracking

The service is an all-in-one hub for request submission, request tracking, and output retrieval. The default landing page displays a table with information about the user's active and past requests, including the creation date, production name, and request status. Clicking on a table entry routes to a dedicated request details page, where users can view

$B^{\pm}$ meson mass example

This example studies the $B^{\pm} \to J/\psi(\to \mu^{+}\mu^{-})K^{\pm}$ decay using a dedicated stripping selection that isolates the signal. Combining the two muons from the J/ψ with a charged kaon reconstructs the $B^{\pm}$ candidate, whose invariant mass forms a sharp peak that can be plotted directly from the ntuples.



Production configuration files

Clone all

File name	File size		
info.yaml	804 Bytes		
Btree.py	742 Bytes		
Btree.yaml	1.177 KiB		

Real production results

Download all

File name	File size	Stream	Year	Magnet polarity		
00334560_00000001_1.dvntuple.root	1.835 GiB	Dimuon	2016	Down		
00334560_00000002_1.dvntuple.root	160.467 MiB	Dimuon	2016	Down		
00334564_00000001_1.dvntuple.root	2.011 GiB	Dimuon	2017	Down		
00334565_00000001_1.dvntuple.root	1.921 GiB	Dimuon	2016	Up		
00334566_00000001_1.dvntuple.root	1.933 GiB	Dimuon	2017	Up		

Usage

The following resources illustrate how to work with these ntuples:

- CERN Open Data record catalogues this dataset on the CERN Open Data Portal.
- LHCb Open Data Guide explains how the `StrippingBs2MuMuLinesBu2JPsiKFullDSTLine` nicely isolates the desired signal and shows how you can analyse the ntuples with ROOT or python.
- reana-demo-lhcb-run2-b2jpsik repository contains a runnable containerised workflow for this analysis.

Figure 4: Expanded view of the $B^{\pm}$ meson mass example using the $B^{\pm} \to J/\psi K^{\pm}$ decay channel.

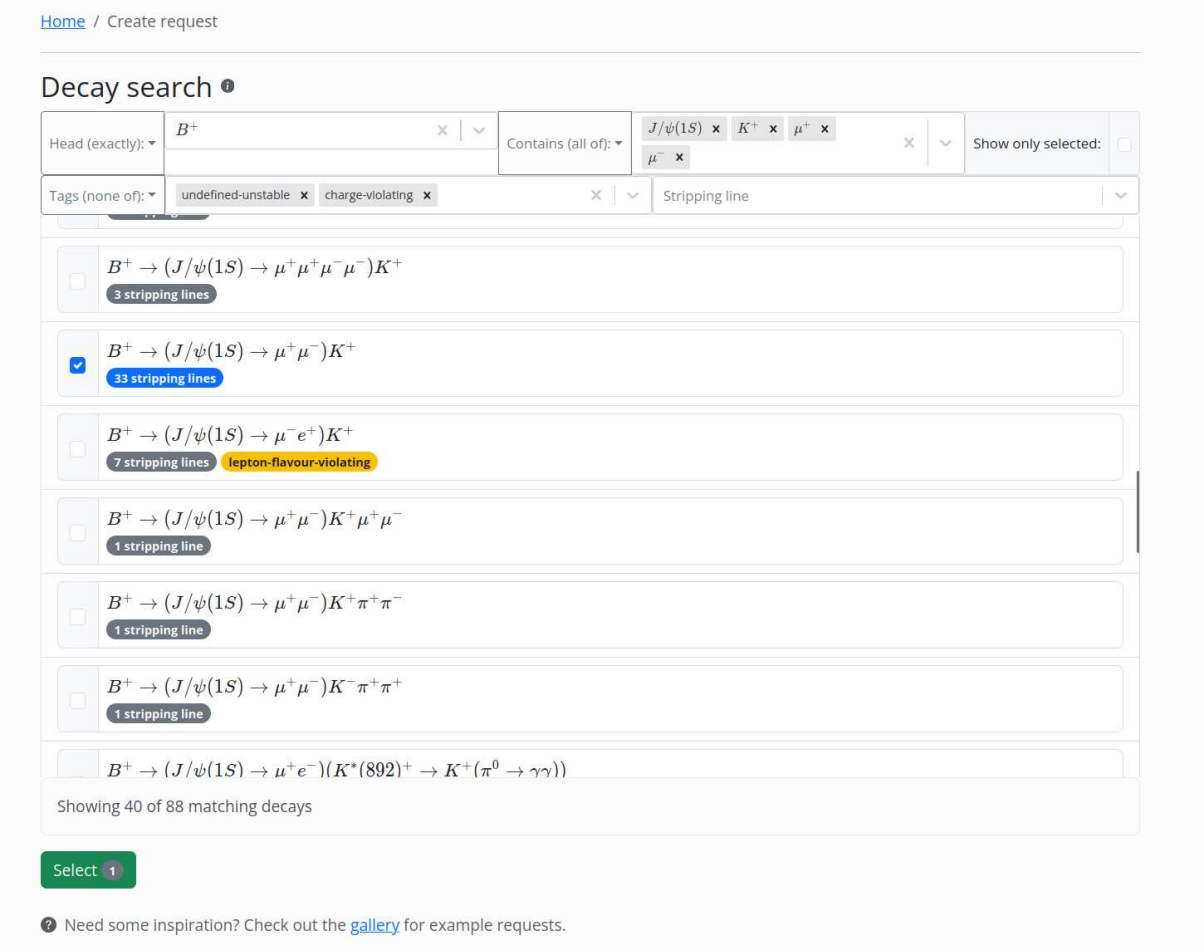


Figure 5: Physics object selection step of the request configuration process.

the full status of their request, the corresponding production configuration files, output ntuples from both the test and full production stages, and a knowledge graph describing many of the variables saved to the ntuple. An example of the request details page can be seen in Figure 8, with the corresponding knowledge graph shown in Figure 9.

The output files are categorized according to the datasets from which they were produced, allowing users to easily identify which data-taking year, magnet polarity, and stream a given ntuple corresponds to. From the request details page, users can inspect ROOT files directly in the browser using JSROOT [35], displaying histograms of data variables. They can also retrieve the produced ntuples either through direct download from the web interface or via supported access protocols such as XRootD [36].

The same page also provides a request-specific communication channel between the user and the LHCb Open Data team. This allows users to raise questions or report concerns directly in the context of a given request, enabling the team to provide targeted, on-demand support at the request level. These messages are displayed in the request database, allowing the Open Data team to respond quickly and keep track of request-specific issues. In addition, users are notified by email for any communication from the LHCb Open Data team, and about relevant status changes, for example when GitLab labels associated with their request are updated, ensuring that they remain informed throughout the production workflow.

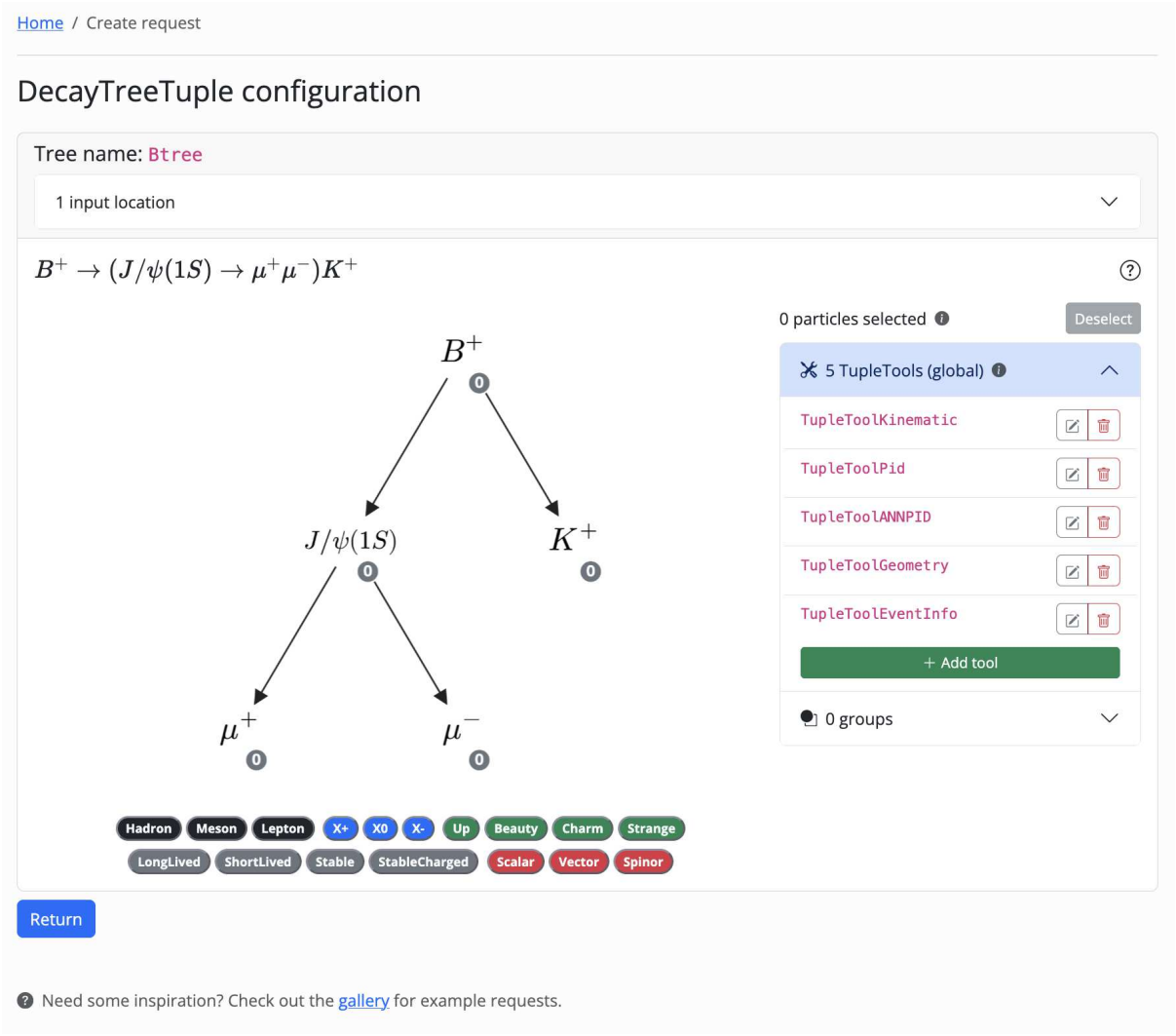


Figure 6: Custom configuration of columns to be added to the ntuple during the request configuration process.

3.4 Usage Uptake

Since the public release of the LHCb Ntupling Service on February 22, 2026, the number of submitted requests has steadily increased. Although the service has only been publicly available for about six months at the time of writing, we have accumulated roughly 100 user requests.

The size of the produced outputs varies substantially between requests, depending on the selected datasets, data-taking years, and requested analysis variables. Small requests may produce output ntuples of only a few hundred MB, while larger requests can result in several hundred GB of output data. For the first 100 requests, the total produced output amounted to approximately 10 TB, corresponding to an average output size of about 100 GB per request.

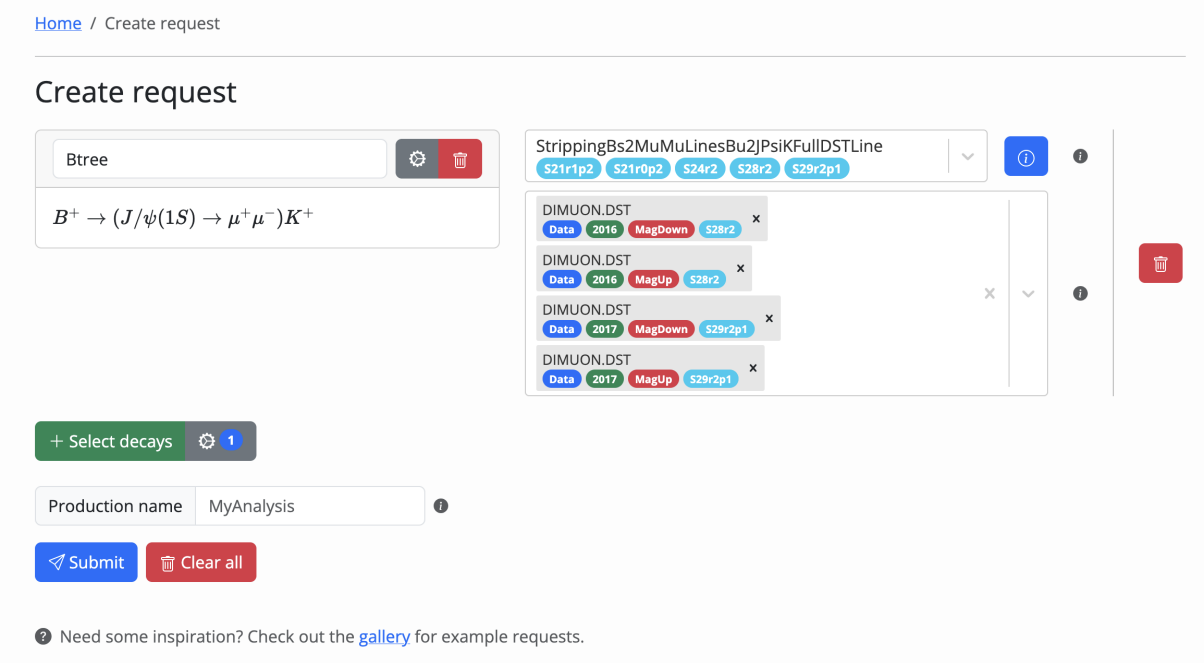


Figure 7: Specification of an LHCb dataset during the request configuration process.

3.5 User Feedback

As part of the request submission procedure, users are asked to briefly describe the intended use of the requested LHCb open data. These descriptions provide useful insight into motivations for accessing ntuples through the LHCb Ntupling Service – they are most relevant for data popularity studies.

The submitted use cases cover a broad range of applications. Some users request data for physics studies, including CP-asymmetry measurements, searches for exotic hadrons, and studies of detector resolution effects. Others use the service to investigate statistical and computational methods, such as fitting algorithms or resolution models, in a realistic HEP environment.

The service is also used in educational contexts. Examples include high-school-level outreach activities, university teaching, and self-guided exercises for learning common analysis tools such as ROOT. In addition, the service provides a lightweight entry point to LHCb data even within the collaboration, for example when training new students before they interact with the full (complexity of the) internal data-processing infrastructure.

The request-specific communication channel has proven useful for understanding user needs and for providing targeted support. Feedback received through these conversations is generally very positive. The ability to ask questions directly in the context of a given request is particularly appreciated. The repeated use of the service by several users after an initial request further suggests that the workflow is useful beyond isolated test cases. At the same time, this feedback is valuable for the development of the service itself, since recurring questions and difficulties help identify where the interface, documentation, or analysis examples can be improved.

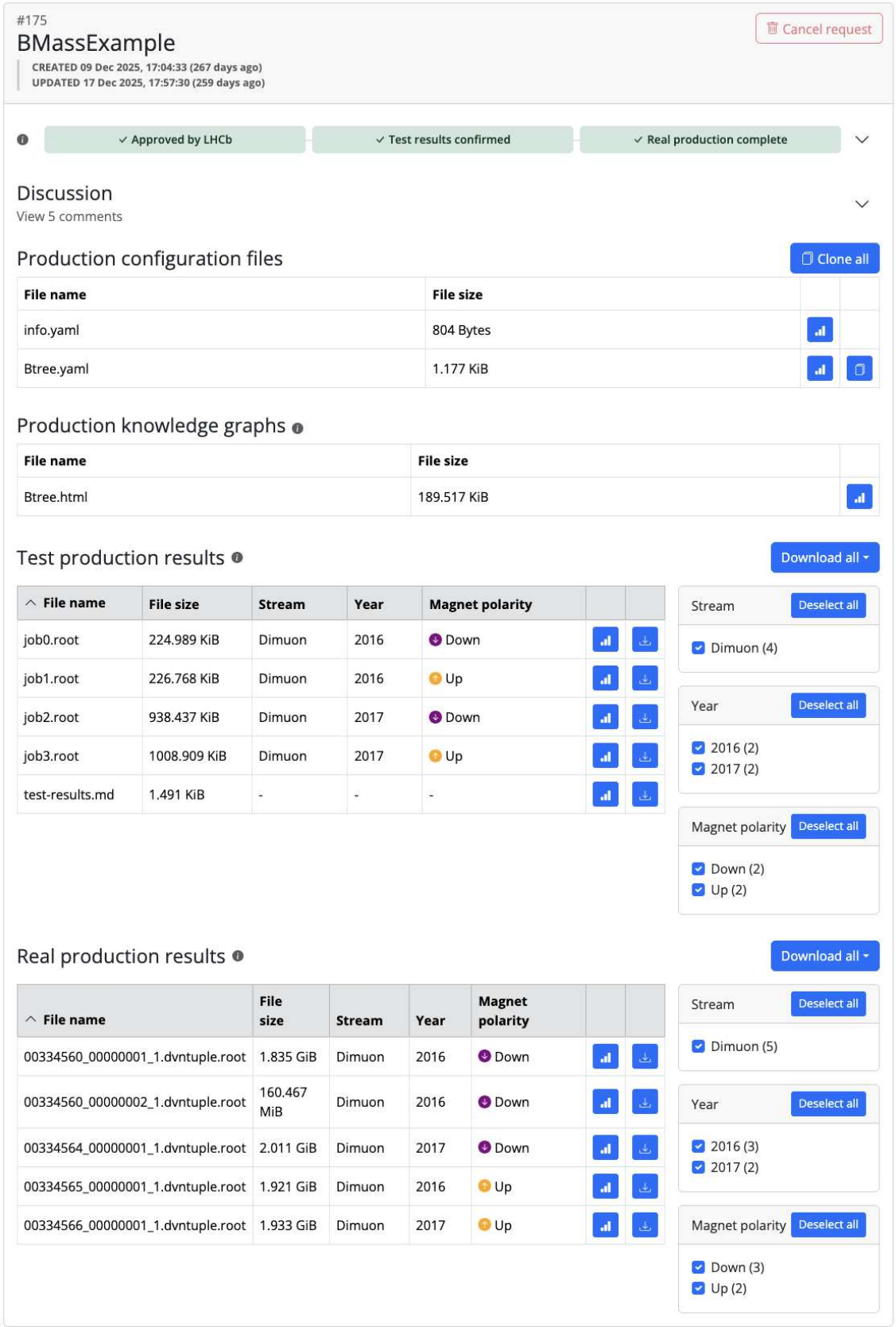


Figure 8: Example of the page users interact with to view their requests.

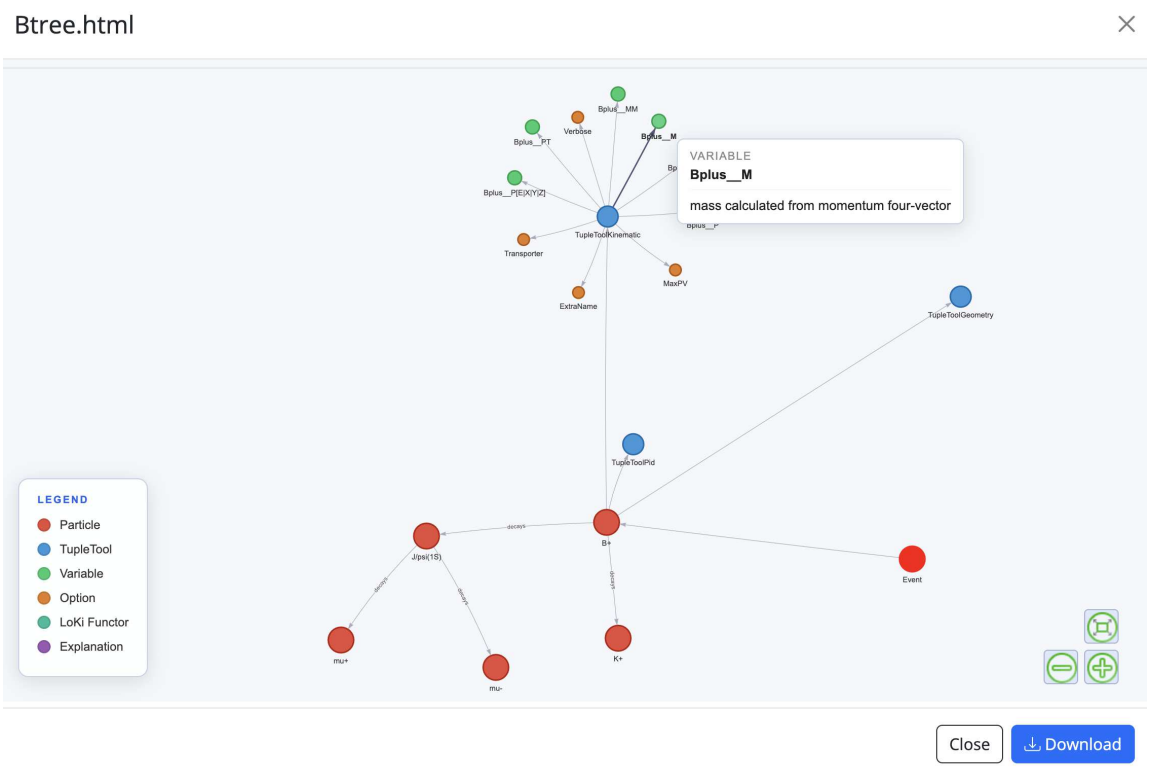


Figure 9: Example of the knowledge graph corresponding to Figure 8, allowing for visualization of available variables in the ntuple.

4 CERN Open Data Portal Integration

The CERN Open Data Portal and the LHCb Ntupling Service operate as separate entities. The CERN Open Data Portal is designed to permanently host static datasets, and it serves as a central hub for the LHC experiments and beyond to host open data. The needs of the LHCb Ntupling Service are fundamentally different, as ntuple creation is done dynamically, and resulting ntuples are not stored permanently unless they are promoted to permanent records by the LHCb Open Data team. However, there are still opportunities for integration between the CERN Open Data Portal and the LHCb Ntupling Service. For example, Section 4.1 discusses the feature of promotion of service requests to permanent records on the CERN ODP, helping to facilitate FAIR principles [10] for data output from the service. In turn, the LHCb Ntupling Service is advertised directly on the CERN Open Data Portal on both the "Getting started with LHCb Open Data" and "About LHCb" pages to promote findability and accessibility of the web service.

4.1 Promotion to Permanent Records

Requests on the service can be promoted to permanent records on the CERN Open Data Portal; a DOI is attributed to the record and corresponding ntuples, and instructions for citation of the data are provided. A procedure is in development to aggregate the necessary metadata and produce the Open Data Portal record in the appropriate `JSON` schema during the request life cycle. Users will be able to request promotion of their ntuples to

a permanent dataset on the CERN Open Data Portal, which will ultimately need to be approved by the LHCb Open Data team. A permanent record for the $B^{\pm}$ meson mass example discussed in Section 3 is shown in Figure 10. As such records accumulate, the gallery of examples advertised on the service will grow, which will directly cross link the CERN Open Data Portal records on the service and provide the option to work with these ntuples rather than producing new ones. This will lead to a favorable trade off of the compute on the WLCG for the relatively light footprint of storing ntuples on the CERN Open Data Portal.

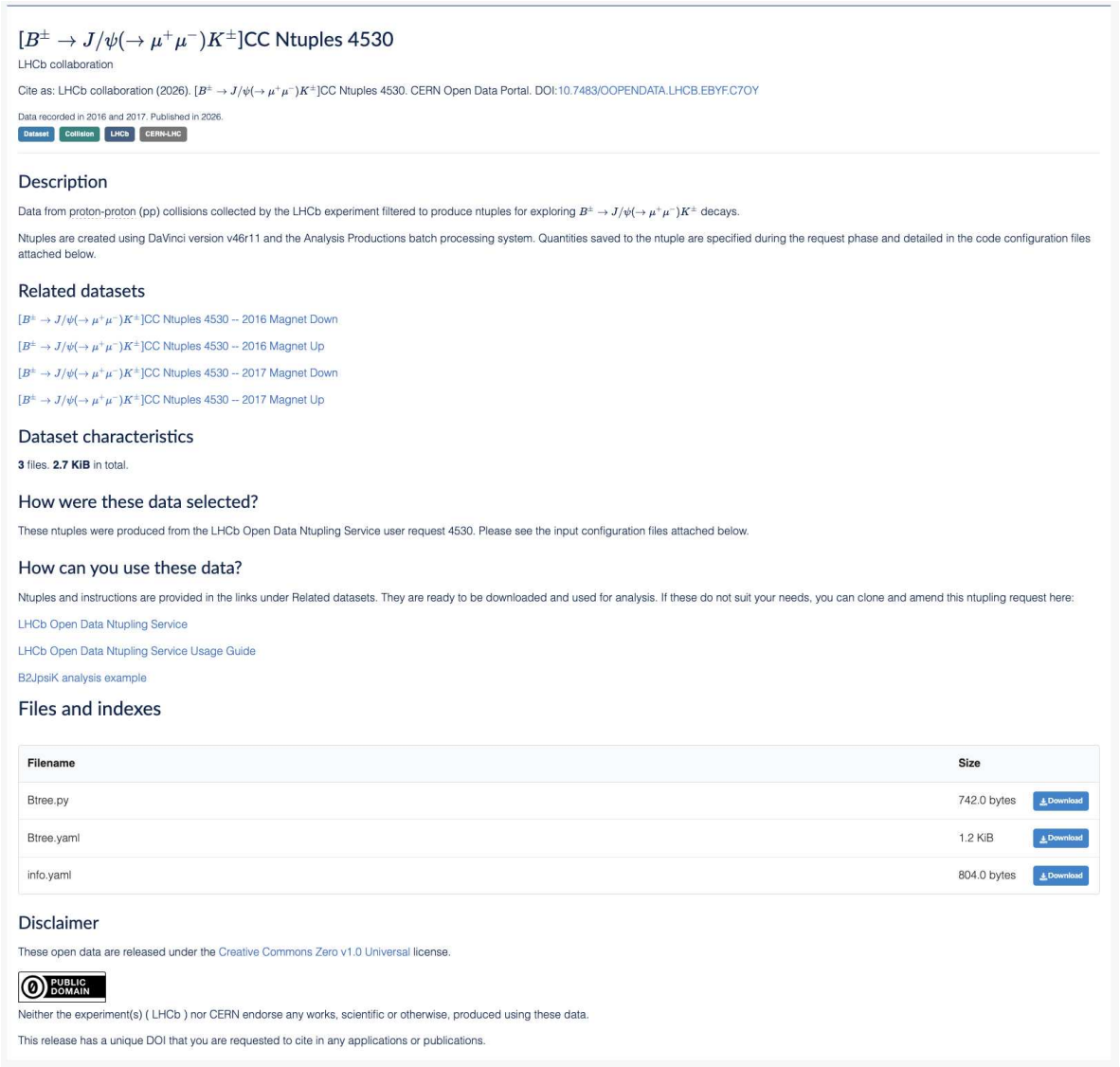


Figure 10: Example of a record on the CERN Open Data Portal resulting from a request from the LHCb Ntupling Service.

4.2 Storage Considerations

As mentioned previously, ntuples produced by the LHCb Ntupling Service are not stored permanently by default. Instead, users are prompted to download their resulting ntuples

onto local data volumes. The time window before deletion of the produced ntuples on the service side is still being decided, and will be informed based on average accumulation of data volume. On the other hand, requests promoted to permanent records on the Open Data Portal will have ntuples stored indefinitely for future use. Both the temporary and permanent data stored from the service must be taken into account relative to the total data volume allocated to the LHCb open data project. At the time of writing, less than 10 TB have been accumulated from LHCb Ntupling Service output, which is only a small fraction of the data volume allocated to LHCb for open data activities. The current rate of data accumulation should give ample time to develop an appropriate solution for temporary storage cleanup, including informing users well in advance before any data deletion.

5 Future Outlook

The LHCb collaboration has now provided open access to ∼100% of the Level-3 proton-proton (pp) collision data collected during LHC's Run 1 (2011-2012) and over 50% of the Level-3 pp collision data collected during Run 2 (2015-2018), in accordance with LHCb's implementation of the CERN Open Data Policy [9]. This outlines LHCb's commitment to making its data publicly accessible: 50% of the data from a given data-taking campaign is to be released five years after its completion, followed by the full dataset after ten years. Level-3 data in this context refers to reconstructed data suitable for research purposes, as defined by the Data Preservation in High Energy Physics (DPHEP) working group [37]. The current implementation in the LHCb Ntupling Service includes the 2011 and 2012 data for Run 1, and the 2015-2017 data for Run 2, excluding the 2018 data. This accounts for roughly all of the integrated luminosity for Run 1 and over half of the integrated luminosity for Run 2. In 2028 it will be 10 years out from Run 2 data taking, prompting the need to provide access to the 2018 data in the service at this time. This will be a simple operation by design, leading to no foreseen delays in providing open access to the full Run 2 LHCb dataset by 2028.

In a similar manner, releases of data collected in other collision systems (such as pPb, Pbp, and PbPb in collider mode, and collisions of protons or lead on noble gas species in fixed target mode with the System for Measuring Overlap with Gas (SMOG)) [38, 39] are achievable with relative ease. Future developments will proceed based on the resources available to the LHCb Open Data project.

No simulation samples are available for request with the LHCb Ntupling Service at the time of writing. However, the LHCb simulation stack is open source, and a documented procedure for generating Monte Carlo (MC) and running it through the LHCb detector simulation as an external user is in development. This would require users to provide their own computing resources in order to generate and process the simulation samples. Other means for providing access to previously prepared simulation samples where only the ntupling step is required are being investigated in parallel. In a similar manner, there is an effort to provide access to calibration samples and internally available tools for calculating track, PID, and trigger efficiencies, but the timescale for achieving this is not yet clear. While this currently presents limitations for publication-grade analyses, these limitations can be overcome with the aforementioned future developments, and the LHCb Open Data team will do what they can to help users in overcoming these hurdles in the

meantime. All developments in this direction will proceed according to resources available to the LHCb Open Data project.

Releases for Run 3 and beyond will take considerable effort, as the infrastructure for ntupling at LHCb has been completely reworked (and improved) [40]. This means that the configuration files derived from the wizard and corresponding internal parsing package described in Ref. [11] would need a major overhaul and separately managed logic for legacy (Run 1 and Run 2) productions and recent and future (Run 3 and beyond) productions. Such an endeavor remains to be discussed within the LHCb collaboration.

6 Summary

The web service presented in this paper, the LHCb Ntupling Service, constitutes a new paradigm for providing open access to data collected in high energy physics experiments. In conjunction with the LHCb Ntuple Wizard, this innovative framework provides access to both Run 1, and for the first time, Run 2 *pp* data collected by the LHCb experiment, amounting to over 4 PB of data to explore. Future releases of 2018 Run 2 data, as well as Run 1 and 2 data with proton-ion or ion-ion collisions can be accomplished with relative ease by design, though releases for Run 3 and beyond will take considerable effort given the changes to the ntuple production software at LHCb. The service provides flexibility for deep exploration of the collected data while reducing the barrier to entry for new users and maintaining a significantly smaller footprint in terms of data volumes relative to historical release mechanisms. It has a robust architecture connecting the LHCb Ntuple Wizard for specification of request configuration, with the Analysis Productions batch processing system for running production jobs on the WLCG. It also provides a single entry point for users to submit ntuple production jobs, track their requests, communicate with the LHCb Open Data team, and download their resulting output files. Substantial documentation and usage hints are provided on the LHCb Open Data Guide, and are embedded directly into the web service. User requests can be promoted to permanent records on the CERN Open Data Portal upon approval from the LHCb Open Data team, in which case the resulting ntuples are stored permanently with citable DOIs.

The LHCb Ntupling Service was officially released in the first quarter of 2026 and has since seen sustained usage with very positive feedback. A wealth of use cases has already been observed, including research into exotic hadrons, development of fitting algorithms, physics education, and much more. This is a significant milestone for both the LHCb experiment and open science in general, allowing for exploration of the unique data collected by the LHCb experiment without the need to learn any experiment-specific software.

Acknowledgements

We thank our colleagues of the LHCb Data Processing & Analysis (DPA) project for their incredibly valuable discussions and enthusiasm. We thank the LHCb Computing and DPA projects for providing the necessary resources for the LHCb Ntupling Service, through the usage of Analysis Productions ran on the WLCG facilities. We also thank the Open Science team at CERN's Department of Information Technology for their extensive support of the LHCb Ntupling Service and for providing storage space for the

resulting data outputs. C.A.A. and D.S.F. acknowledge support from the National Science Foundation under Award Number 2309922. We further acknowledge support from the national German agencies BMFTR and PUNCH4NFDI (DFG).

References

[1] LHCb collaboration, *LHCb computing: Technical Design Report*, CERN-LHCC-2005-019, 2005.

[2] LHCb collaboration, *LHCb Upgrade Software and Computing*, CERN-LHCC-2018-007, 2018.

[3] CMS collaboration, G. L. Bayatyan *et al.*, *CMS computing: Technical Design Report*, CERN-LHCC-2005-023, CMS-TDR-7, CERN, 2005.

[4] CMS collaboration, G. Petrucciani, A. Rizzi, and C. Vuosalo, *Mini-AOD: A New Analysis Data Format for CMS*, J. Phys. Conf. Ser. **664** (2015) 7, arXiv:1702.04685.

[5] ATLAS collaboration, G. Duckeck *et al.*, *ATLAS computing: Technical Design Report*, CERN-LHCC-2005-022, ATLAS-TDR-017, CERN, 2005.

[6] ATLAS collaboration, G. Aad *et al.*, *Software and computing for Run 3 of the ATLAS experiment at the LHC*, Eur. Phys. J. C **85** (2025) 234, arXiv:2404.06335.

[7] ALICE collaboration, P. Cortese *et al.*, *ALICE computing: Technical Design Report*, CERN-LHCC-2005-018, ALICE-TDR-12, CERN, 2005.

[8] ALICE collaboration, P. Buncic, M. Krzewicki, and P. Vande Vyvre, *Technical Design Report for the Upgrade of the Online-Offline Computing System*, CERN-LHCC-2015-006, ALICE-TDR-019, CERN, 2015.

[9] *CERN Open Data Policy for the LHC Experiments*, CERN-OPEN-2020-013, CERN, Geneva, 2020. doi: 10.17181/CERN.QXNK.8L2G.

[10] M. D. Wilkinson *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship*, Sci. Data **3** (2016) 160018.

[11] C. A. Aidala *et al.*, *Ntuple Wizard: An Application to Access Large-Scale Open Data from LHCb*, Comput. Softw. Big Sci. **7** (2023) 6, arXiv:2302.14235.

[12] A. Abdelmotteleb *et al.*, *The LHCb Sprucing and Analysis Productions*, Comput. Softw. Big Sci. **9** (2025) 15, arXiv:2506.20309.

[13] LHCb collaboration, *LHCb Ntupling Service*, https://opendata-lhcb-ntupling-service.app.cern.ch/, 2026.

[14] C. Aidala *et al.*, *LHCb Open Data Ntupling Service: On-demand production and publishing of custom LHCb Open Data*, EPJ Web Conf. **337** (2025) 01231, arXiv:2504.00610.

[15] LHCb collaboration, *DaVinci*, https://gitlab.cern.ch/lhcb/DaVinci.

[16] R. Brun and F. Rademakers, *ROOT – An object oriented data analysis framework*, Nucl. Instrum. Meth. **A389** (1997) 81.

[17] I. Antcheva *et al.*, *ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization*, Comput. Phys. Commun. **180** (2009) 2499, arXiv:1508.07749.

[18] CERN, *CERN Open Data Portal*, https://opendata.cern.ch/.

[19] CMS collaboration, K. Ehatäht, *NANOAOD: a new compact event data format in CMS*, EPJ Web Conf. **245** (2020) 06002.

[20] Schaarschmidt, Jana *et al.*, *PHYSLITE - A new reduced common data format for ATLAS*, EPJ Web of Conf. **295** (2024) 06017.

[21] ATLAS collaboration, M. V. Albornoz, *The First Release of ATLAS Open Data for Research*, PoS **ICHEP2024** (2025) 1172.

[22] K. Lassila-Perini, C. Lange, E. C. Jarrin, and M. Bellis, *Using CMS Open Data in research – challenges and directions*, EPJ Web Conf. **251** (2021) 01004, arXiv:2106.05726.

[23] Meta Open Source, *React: A JavaScript library for building user interfaces*, https://react.dev/.

[24] GitLab Inc., *GitLab*, https://about.gitlab.com/.

[25] W. Reese, *Nginx: The high-performance web server and reverse proxy*, Linux Journal **2008** (2008) 2.

[26] R. Cox, R. Griesemer, and R. Pike, *The Go Programming Language and Environment*, Communications of the ACM **65** (2022) 70.

[27] LHCb collaboration, *LHCb Ntuple Wizard*, https://www.npmjs.com/package/lhcb-ntuple-wizard, 2026.

[28] npm, *npm*, https://www.npmjs.com/.

[29] LHCb collaboration, *LHCb Ntuple Wizard*, https://gitlab.cern.ch/lhcb-dpa/wp6-analysis-preservation-and-open-data/lhcb-ntuple-wizard-frontend, 2026.

[30] A. J. Peters and D. C. van der Ster, *Evaluating CephFS Performance vs. Cost on High-Density Commodity Disk Servers*, Comput. Softw. Big Sci. **5** (2021) 25.

[31] A. Peters, E. Sindrilaru, and G. Adde, *EOS as the present and future solution for data storage at CERN*, J. Phys. Conf. Ser. **664** (2015) 042042.

[32] LHCb collaboration, *LHCb Open Data Guide*, https://lhcb-opendata-guide.web.cern.ch/, 2026.

[33] T. Šimko *et al.*, *REANA: A System for Reusable Research Data Analyses*, EPJ Web Conf. **214** (2019) 06034.

[34] N. A. Grieser *et al.*, *The LHCb Stripping Project: Sustainable Legacy Data Processing for High-Energy Physics*, Comput. Softw. Big Sci. **9** (2025) 21, arXiv:2509.05294.

[35] B. Bellenot and S. Linev, *JavaScript ROOT*, J. Phys. Conf. Ser. **664** (2015) 062033.

[36] G. Amadio *et al.*, *XRootD Client: a robust technology for LHC Run-3 and beyond*, EPJ Web Conf. **295** (2024) 01056.

[37] DPHEP Study Group, 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.

[38] G. Graziani *et al.*, *LHCb Physics Motivations for the 2016 Heavy-ion LHC run*, LHCb-PUB-2016-011, 2016.

[39] A. Bursche *et al.*, *Physics opportunities with the fixed-target program of the LHCb experiment using an unpolarized gas target*, LHCb-PUB-2018-015, 2018.

[40] A. Mathad *et al.*, *FunTuple: A new N-tuple component for offline data processing at the LHCb experiment*, Comput. Softw. Big Sci. **8** (2024) 6, arXiv:2310.02433.