

CERN Open Data

CS3 2024

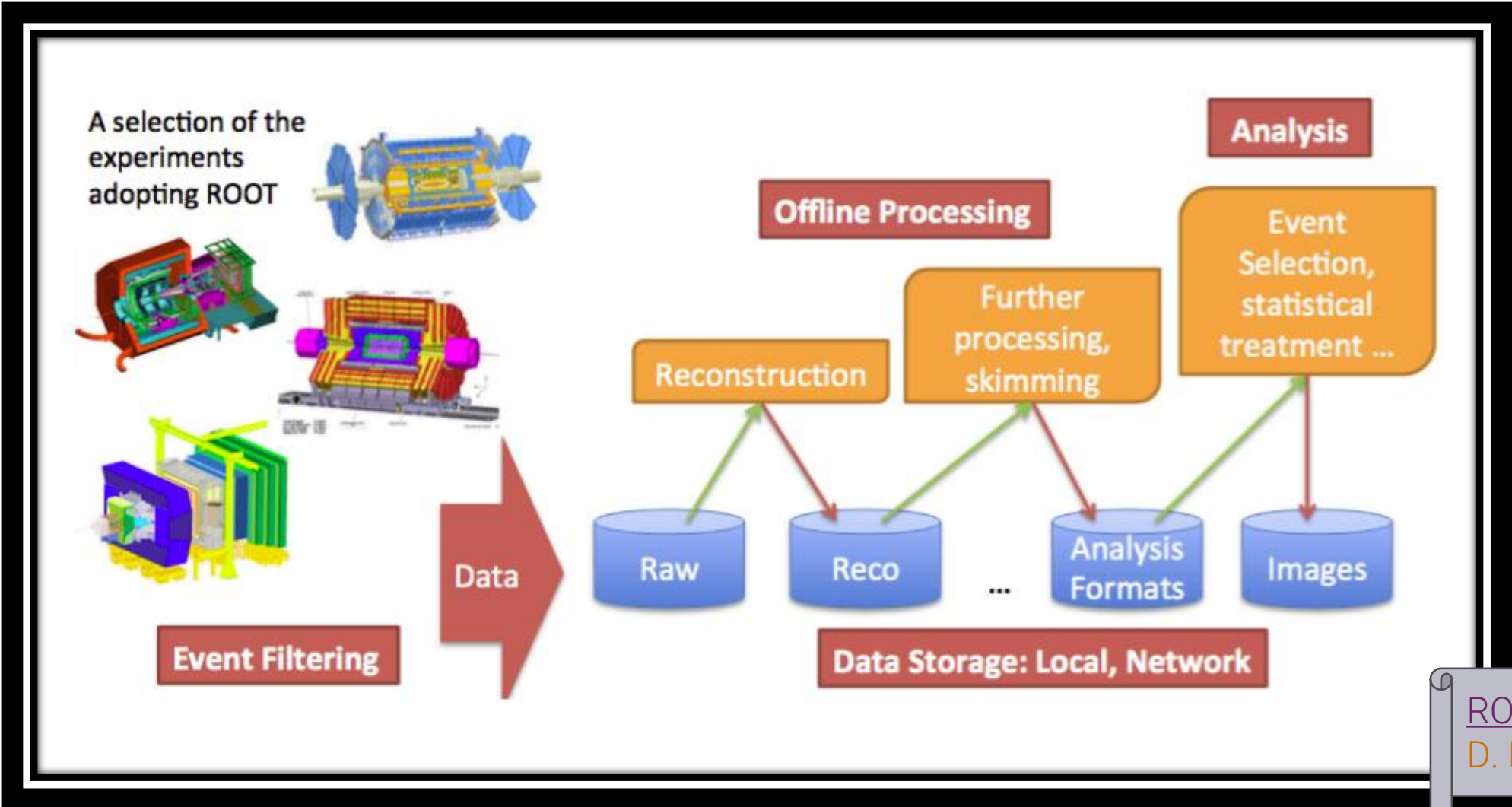
Pablo Saiz

12 Mar 2024

Large Hadron Collider



High Energy Physics data levels



ROOT Tutorial
D. Krücker et al.

LHC data preservation and open access policies

ALICE data preservation strategy

Sunday, October 6, 2013

The data harvested by the ALICE Experiment up to now and to be harvested in the future constitute the return of investment in human and financial resources by the international community. These data embed unique scientific information for the in depth understanding of the profound nature and origin of matter. Because of their uniqueness, long term preservation must be an essential objective of the data processing framework and will be the foundation of the ALICE Collaboration legacy to the scientific community as well as to the general public. These considerations call for a detailed assessment of the ALICE data preservation strategy and policy. Documentation, long term preservation at various levels of abstraction, data access and analysis policy and software availability constitute the key elements of such a data preservation strategy allowing future collaboration, the wider scientific community and the general public to analyze data for educational purpose and for overall advancement of the published results. The present document describes the basic principles that will guide the induction addressed by the ALICE data preservation policy.

ALICE data formats

The level of abstraction of ALICE data increases at every step of the data processing chain starting from basic raw data delivered by the detectors of the experiment, evolving into physics analysis-ready data and ending with physics data suitable for publication. At each stage of the data processing ancillary meta-data, such as calibration, alignment and running condition parameters are included to transform raw detector information into physics information, free of detector biases. The various ALICE data formats are classified as follows:

- 1) Raw data embedding the signals delivered by the detectors along with the associated status data containing various information on the running conditions constituting the primary information collected by the ALICE experiment. They provide the input of the reconstruction algorithms, together with the calibration data stored in a dedicated database;
- 2) Monte-Carlo data, including data at the event generator level (MC truth) and data mimicking the raw data format (digits), anchored to real data reproducing the running conditions;
- 3) Event Summary Data (ESD) produced by the reconstruction algorithms, for both Monte Carlo and raw data. The ESD events provide calibrated tracks in a generic format, but also additional detector specific information allowing a full physics analysis;
- 4) General purpose Analysis Object Data (AOD), derived from ESD data. The AOD data format contains a simplified event model with few additional high level detector specific parameters;
- 5) Custom analysis objects, used standalone or together with the general purpose AOD for specific analysis;
- 6) Published physics results and highly abstracted data resulting from the analysis.

These different formats of the ALICE data lead to a specific schema for data preservation. While formats can change with time, the collaboration provides software releases suitable to read and process any format, or alternatively to migrate data from one format to another. Since processed data can exist in several versions, only the version used for the final publication of the results is considered as a candidate for data preservation.

The ALICE Computing Model includes the provision for permanent storage of two copies of the raw data. They are not presently being considered for open access, but they can be reprocessed at any time by members of the ALICE collaboration upon approval by the ALICE Physics Board. The original datasets used to produce published results, together with the adequate software version (framework and macros) are subject to long-term preservation.

ATLAS Data Access Policy

May 21st 2014

Introduction

ATLAS has fully supported the principle of open access in its publication policy. This document outlines the policy of ATLAS regarding open access to data at different levels as described in the CERN P12 model. The main objective is to make the data available in a suitable way to people external to the ATLAS collaboration.

The ATLAS policy for data preservation is described in a separate document. The collaboration's need to preserve data for its own use shares some requirements with making them open access. To support open access to data additional resources will be required to develop and support the tools to make the data available.

Policies for Different Data Levels

Open access to ATLAS data by people outside the collaboration can be considered at four levels of increasing complexity, listed below, with associated conditions, see Ref. [3]. This policy pertains to collision physics data (i.e. that are stored online and intended for physics analysis) and the necessary associated metadata, along with associated simulated datasets and tools allowing to produce new simulated datasets based on an adequate simulation of the ATLAS detector.

Level-1. Published results

All scientific output published in journals, and preliminary results are made available in Conference Notes. All are openly available, without restriction on use by external parties beyond copyright law and the standard conditions agreed by CERN.

Data associated with journal publications are also made available: tables and data from plots (e.g. cross section values, likelihood profiles, selection efficiencies, cross section limits, ...) are stored in appropriate repositories such as HEPDATA[2]. ATLAS also strives to make additional material related to the paper available that allows a reinterpretation of the data in the context of new theoretical models. For example, an extended interpretation of the analysis is often provided for measurements in the framework of RISE[3]. For searches information on signal acceptance is also made available to allow reinterpretation of these searches in the context of models developed by theorists after the publication. ATLAS is also exploring how to provide the capability for reinterpretation of searches in the future via a service such as RECAST[4]. RECAST allows theorists to evaluate the sensitivity of a published analysis to a new model they have developed by submitting their model to ATLAS.

Level-2. Outreach and Education

ATLAS recognizes the vital role of outreach and education, and participates in and encourages outreach and education activities, and makes selected data available for them. Typically a fraction of the complete ATLAS data set is used, selected to provide a rich sample of events with interesting physics signatures but not adequate for a publication of a physics result. The data are provided in compressed, portable and self-contained formats for

CMS data preservation, re-use and open access policy

Version 1.2
Approved by the CMS Collaboration Board 20th April 2018

CMS data are unique and are the result of vast and long-term moral, human and financial investment by the international community. There is unique scientific opportunity in re-using these data, at different levels of abstraction and at different points in time. This opportunity calls for our collective responsibility and poses unprecedented challenges, as no data sample of the complexity and value has ever been preserved or made available for later re-use.

The CMS collaboration is committed to preserve its data, at different levels of complexity, and to allow their re-use by a wide community including: collaboration members long after the data are taken, experimental and theoretical HEP scientists who were not members of the collaboration, educational and outreach initiatives, and citizen scientists in the general public.

CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. To that extent, CMS will provide open access to its data after a suitable but relatively short embargo period, allowing CMS collaborators to fully exploit their scientific potential.

This policy describes the CMS principles of data preservation, re-use and open access, as well as the relevant access in all these tasks and their roles and responsibilities. CMS understands that in order to fully exploit all these re-use opportunities, immediate and continued resources are needed. The level of support that CMS will be able to provide to external users depends on the available funding. This policy addresses the moral responsibility of CMS for its data, as well as the increasing concern of funding agencies worldwide and the civil society for the preservation and re-use of scientific data.

Notwithstanding the long-term perspective of the LHC programme, the time for action is now: lower-energy and lower-luminosity LHC runs at centre-of-mass energies of 8.8, 2.36, 2.76, 7 and 8 TeV may never be repeated, and their preservation and preparation for later re-use, has to be addressed urgently. Meeting this challenge is a unique way to stress test and evaluate the entire preservation, re-use and open access concept for the CMS data.

CMS data take many forms. Starting from either raw experimental or simulated data through to reconstructed data and the datasets of higher abstraction generated by a physics workflow, and finally all the way to data represented in scientific publications. Each of these layers has the potential to afford different opportunities for long-term re-use and poses different challenges for preservation. Data represented in publications can already be preserved by building on the existing practices of the Collaboration (e.g. open access publishing) and using third-party platforms (e.g. INSPIRE¹), simply expanding the concept of publication to include additional data sets of a high level of abstraction. At the other extreme of the spectrum, closer to the raw data, different challenges appear which imply a paradigm shift from in-depth documenting and archiving of analyses during the publication process, to a preservation of reconstruction and simulation software packages with all their dependencies.

¹ INSPIRE: <http://inspirehep.net>. ² HEPDATA: <http://hepdata.cern.ch/>. ³ RISE: <http://rise.cern.ch/>. ⁴ RECAST: <http://recoast.cern.ch/>.

LHCb Data Access Policy
LHCb Public Note
Version 1.0
Date: 20th April 2014

Reference: LHCb-PUB-2014-001
Revision: 1
Last modified: 20th April 2014

Abstract

This document contains the LHCb Data Access Policy. This was adopted at the Collaboration Board meeting of 27th Feb 2013.

Data Access Policy for LHCb

1. Data preservation is fundamentally important for the collaboration itself, regardless of any external requirements. This is to enable collaboration members to access data for many years after it was taken and requires a consistent set of the data, associated software, metadata and conditions and documentation to be preserved. LHCb will seek to develop such a data preservation capability as soon as practical. We will need to identify additional resources for this.
2. LHCb supports the principle of open access. In principle we can envisage providing some such open access based upon the work needed internally for data preservation (point 1 above).
3. LHCb is extremely resource limited at present. Therefore whilst this policy expresses a spirit of intent, we cannot commit to implementation of any capability on any specific timescale. Specifically in respect of open access we will not be able to undertake any significant development to support this without injection of additional resources.
4. Overall the collaboration expects to follow the guidelines being developed by CERN and the LHC experiments (initially on these matters, after appropriate approval by the LHCb Collaboration Board).
5. Open access to its data by people outside the collaboration can be considered at four levels of increasing complexity, listed below, with associated conditions. [note: these "levels" 1-4 are those arising in the DUNE model, and are often referred to as such by all the experiments]. In this first iteration, this policy only pertains to collision physics data (i.e. that are online and destined for physics analysis).
6. This policy is adopted by LHCb in good faith according to the spirit of the principles. The collaboration reserves the right to review the policy at any time in the light of experience including, but not limited to, the policy being found to be inadequate in the light of actual requests or any other unintended consequences arising.

page 1



Restricted data → Embargo period (~5 years) → Open data

CERN Open Data Portal

- **Repository of data for the general public**

- Launched in November 2014
- Using Invenio Framework



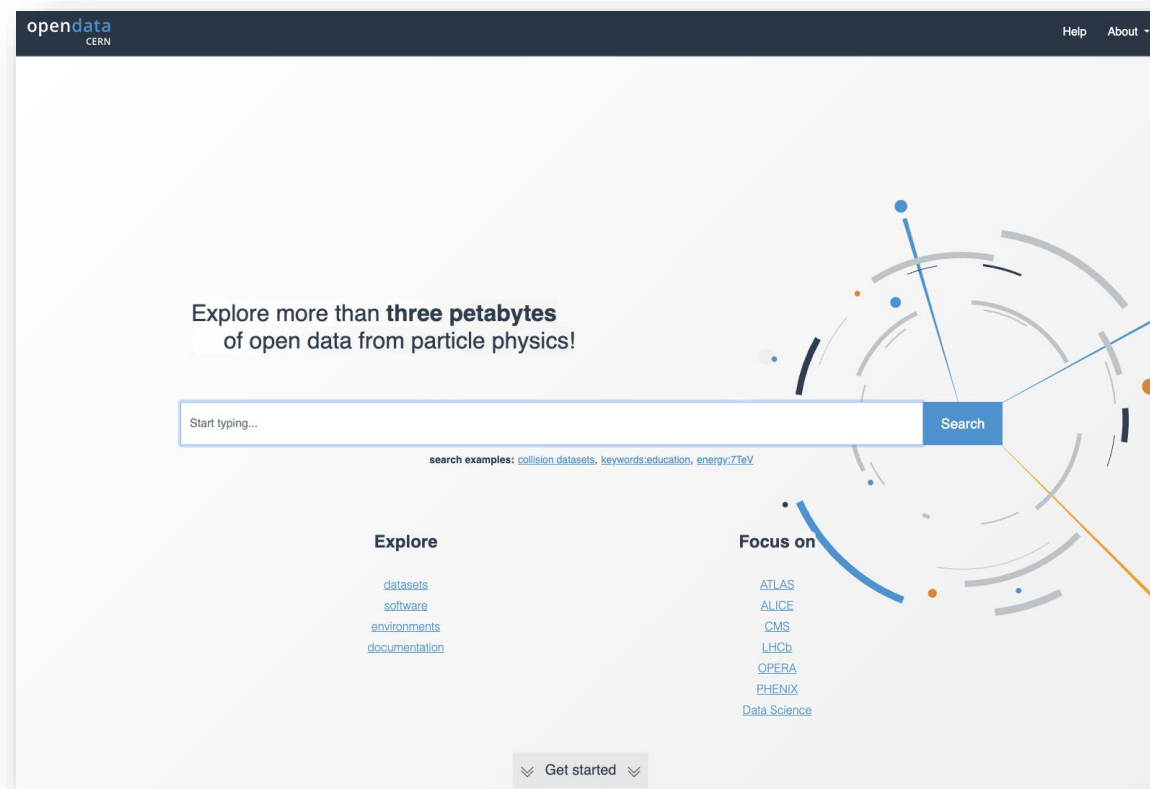
- **Plenty of content**

- Dataset
 - Collision, simulated and derived
- Documentation
 - Glossary, tutorials, configuration, examples
- Software
 - Frameworks, virtual machines, containers

- **Current size (Mar 2024)**

- > 24.000 records
- > 1.900.000 files
- > 4.5 PB

<http://opendata.cern>



Developed by CERN in collaboration with Experiments



CERN Open Data

Higgs-to-four-lepton analysis example using 2011-2012 data

Authors: Jomhari, Nur Zulaha, Getser, Achim, Bin Anuar, Aliq Alzaidin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.2483/OPENDATA.CMS.J08B.R04Z

Description

This research level example is a strongly simplified reimplementation of parts of the original CMS Higgs to four lepton analysis published in Phys.Lett. B716 (2012) 30-61, arXiv:1207.7235.

The published reference plot which is being approximated in this example is https://inspirehep.net/record/1124338/files/H4L_mss_3.png. Other Higgs final states (e.g. Higgs to two photons), which were also part of the same CMS paper and strongly contributed to the Higgs boson discovery, are not covered by this example.

The example consists of different levels of complexity. The highest level of this example addresses users who feel they have at least some minimal understanding of the physics of this paper and of the measurement of this reference plot which can be executed via (reusable) educational level with the linux openstack environment.

Use with

The example uses publication due to but not identical in many later CMT: /DoubleElectron/ /DoubleMu/Run2

Run research-grade analysis examples

opendata CERN

Search

Need HELP?

SPY WebGL

Multi Events Run: 146436/Event: 90755600 (9 of 26)

Detector

- Pixel Barrel
- Pixel Endcap (+)
- Pixel Endcap (-)
- Tracker Inner Barrel
- Tracker Outer Barrel
- Tracker Inner Detector (+)
- Tracker Inner Detector (-)
- Tracker Endcap (+)

CMS Experiment at the LHC, CERN
Data recorded: 2010-Sep-22 21:26:57:387024 GMT
Run / Event / LS: 146436 / 90755600 / 322

Click on a name under "Provenance", "Tracking", "ECAL", "HCAL", "Muon", and "Physics" to view contents in table

opendata CERN

Search

Need HELP?

146436 events with invariant mass between 2-5 GeV

Select one or more parameters:

E1 pT1 eta1 pT1 Q1 E2 pT2 eta2 pT2 Q2 M

Log X Log Y Set bin width Set Undo selection(s)

E1

The total energy of the first lepton [GeV]

Log X Log Y Set bin width Set Undo selection(s)

eta1

The pseudorapidity of the first muon

Log X Log Y Set bin width Set Undo selection(s)

E2

The total energy of the second lepton [GeV]

Log X Log Y Set bin width Set Undo selection(s)

eta2

The pseudorapidity of the second lepton

Controls:

Click on the "LogX" and "LogY" buttons to transform the axes by log10

Enter a bin width and click "Set" to change the bin width of the histogram (the default is 0.1)

Click on the histogram and move to select a region along the x axis. Click "Undo selection(s)" to return to the original range.

Interactive event display and histograms

Independent research...

INSPIRE HEP

literature

Help Submit Login

Literature Authors Jobs Seminars Conferences More...

Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski (Reed Coll.), Simone Marzani (SUNY, Buffalo), Jesse Thaler (MIT, Cambridge, CTP), Aashish Tripathi (MIT, Cambridge, CTP), Wei Xue (MIT, Cambridge, CTP)
Apr 17, 2017

7 pages
Published in: Phys.Rev.Lett. 119 (2017) 13, 132003
Published: Sep 26, 2017
e-Print: 1704.05066 [hep-ph]
DOI: 10.1103/PhysRevLett.119.132003
Report number: MIT-CTP-4891
View in: ADS Abstract Service

pdf cite claim reference search 63 citations

Citations per year

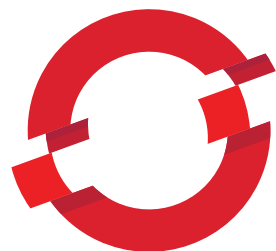
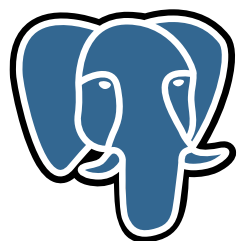
2017 2019 2021 2023

Abstract: (APS)

The splitting function is a universal property of quantum chromodynamics (QCD) which describes how energy is shared between partons. Despite its ubiquitous appearance in many QCD calculations, the splitting function cannot be measured directly, since it always appears multiplied by a collinear singularity factor. Recently, however, a new jet substructure observable was introduced which asymptotes to the splitting function for sufficiently high jet energies. This provides a way to expose the splitting function through jet substructure measurements at the Large Hadron Collider. In this Letter, we use public data released by the CMS experiment to study the two-prong substructure of jets and test the $1 \rightarrow 2$ splitting function of QCD. To our knowledge, this is the first ever physics analysis based on the CMS Open Data.

... that the CMS collaboration cites

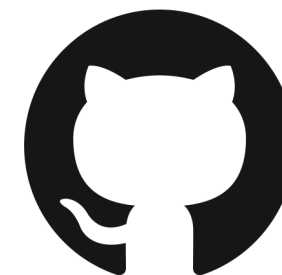
Technologies



OPENSIFT



XRRootD




FAIR guiding principles

- Findable
- Accessible
- Interoperable
- Reusable

<https://www.nature.com/articles/sdata201618>


[Open access](#) | [Published: 15 March 2016](#)

The FAIR Guiding Principles for scientific data management and stewardship

[Mark D. Wilkinson](#), [Michel Dumontier](#), [IJsbrand Jan Aalbersberg](#), [Gabrielle Appleton](#), [Myles Axton](#), [Arie Baak](#), [Niklas Blomberg](#), [Jan-Willem Boiten](#), [Luiz Bonino da Silva Santos](#), [Philip E. Bourne](#), [Jildau Bouwman](#), [Anthony J. Brookes](#), [Tim Clark](#), [Mercè Crosas](#), [Ingrid Dillo](#), [Olivier Dumon](#), [Scott Edmunds](#), [Chris T. Evelo](#), [Richard Finkers](#), [Alejandra Gonzalez-Beltran](#), [Alasdair J.G. Gray](#), [Paul Groth](#), [Carole Goble](#), [Jeffrey S. Grethe](#), ... [Barend Mons](#)  [+ Show authors](#)

[Scientific Data](#) **3**, Article number: 160018 (2016) | [Cite this article](#)

653k Accesses | **6374** Citations | **2138** Altmetric | [Metrics](#)

 An [Addendum](#) to this article was published on 19 March 2019

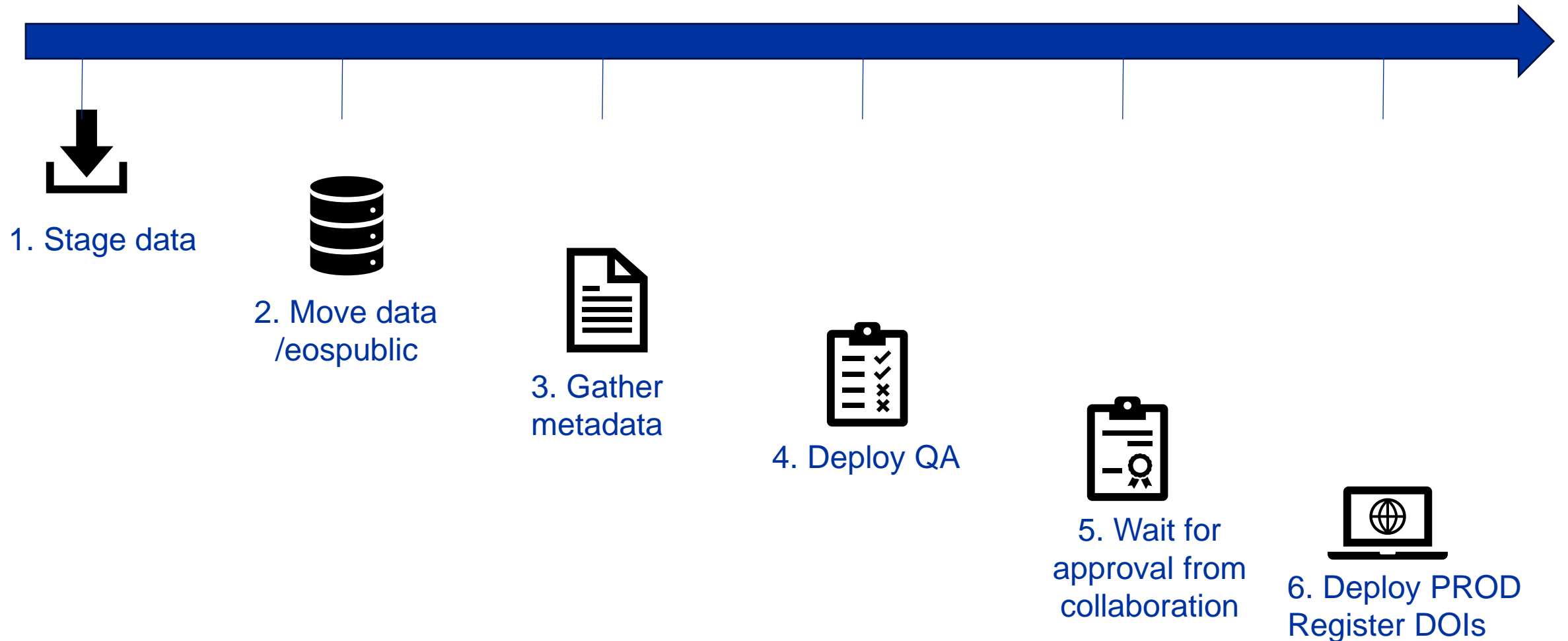
Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

[Digital repositories for FAIR data management](#)
Lars Holm Nielsen



Curating content



CERN Open Data dataset access



File Indexes

Filename	Size	
CMS_HIRun2013_PAMuon_RECO_28Sep2013-v1_10000_file_index.txt	96.5 KiB	List files Download
CMS_HIRun2013_PAMuon_RECO_28Sep2013-v1_20000_file_index.txt	115.7 KiB	List files Download
CMS_HIRun2013_PAMuon_RECO_28Sep2013-v1_20001_file_index.txt	266.0 bytes	List files Download

PAMuon primary dataset in RECO format from the 5.02 TeV proton-Pb run of 2013 (/PAMuon/HIRun2013-28Sep2013-v1/RECO)

ATLAS collaboration (2023). PAMuon primary dataset in RECO format from the 5.02 TeV proton-Pb run of 2013 (/PAMuon/HIRun2013-28Sep2013-v1/RECO). CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.B1W4.HP92

Description

PAMuon primary dataset in RECO format from the 5.02 TeV proton-Pb run of 2013. Reprocessing for run numbers from 210498 to 210658. For run numbers greater than or equal to 210676, please use the related PromptReco-v1 dataset linked below. During the 2013 proton-Pb data taking, the beam direction was reversed between runs 211256 and 211313. The list of validated runs, which must be applied to all analyses, either with the full validation or for an analysis requiring only muons, can be found in Validated runs, full validation. Validated runs, muons only.

Related datasets

The corresponding PromptReco-v1 dataset: /PAMuon/HIRun2013-PromptReco-v1/RECO

Dataset characteristics

15445691 events, 1636 files, 5.0 TiB in total.

Dataset characteristics

15445691 events, 1636 files, 5.0 TiB in total.

List of files

Filename	Size
0002F563-8A2B-E311-8EC3-782BC38F205.root	2.0 GiB

Please note that the file you are going to download (009F5CD5-7D2B-E311-BA17-D4AE528FF49B.root) is **2.1 GiB** big. On an average ADSL connection, it may take several hours to download it.

Moreover, if you use one of the provided **Virtual Machines** to perform your analyses, then you don't need to download datasets manually, because the VM will fetch all the necessary file chunks via the XRootD protocol.

Manual download of files via HTTP is only necessary if you would prefer not to use the XRootD protocol for one reason or another.

Cancel Download

```
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10000
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10001
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10002
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10003
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10004
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10005
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10006
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10007
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10008
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10009
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10010
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10011
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10012
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10013
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10014
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10015
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10016
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10017
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10018
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10019
root://eospublic.cern.ch/eos/opendata/cms/hidata/HIRun2013/PAMuon/RECO/28Sep2013-v1/10020
```



Optimize storage cost

- **Given:**

- Amount of data: 4.5 PB, and increasing
- Long term data preservation
- Accepted latency
- Multiple copies of data (beware: data ownership)

- **Are there any more efficient ways of long term archival?**

- **Possibilities:**

- Computing instead of storage ([LHCb Ntupling service](#))
- Cold Storage. Move part of the data to cheaper storage (Tape, experiment framework)

Cold Storage integration



PAMuon primary dataset in RECO format from the 5.02 TeV proton-Pb run of 2013 (/PAMuon/HIRun2013-28Sep2013-v1/RECO)

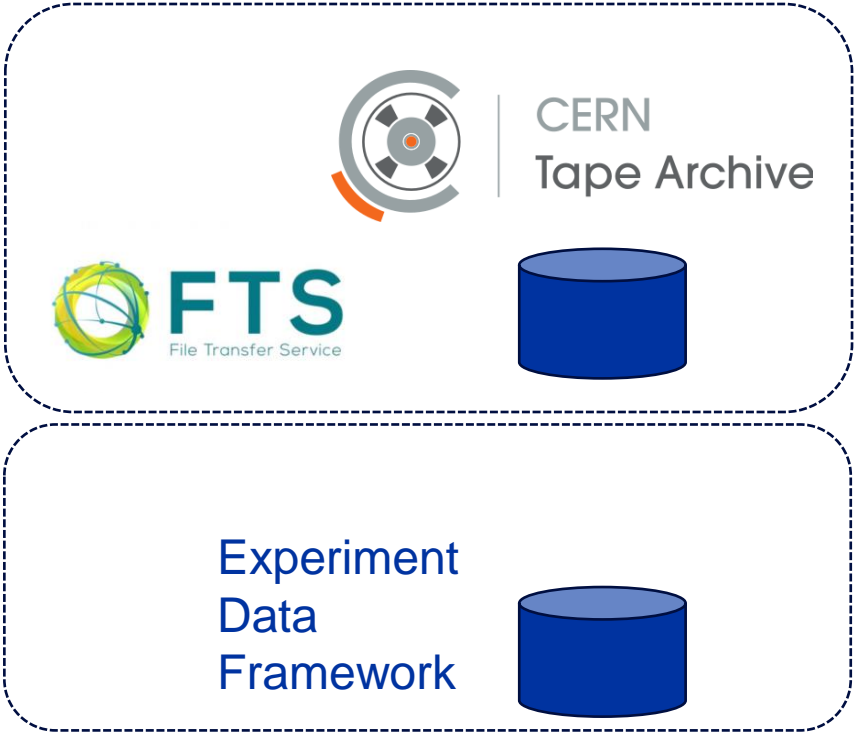
Description
PAMuon primary dataset in RECO format from the 5.02 TeV proton-Pb run of 2013. Reprocessing for run numbers from 210498 to 210658. For run numbers greater than or equal to 210676, please use the related PromptReco-v1 dataset linked below. During the 2013 proton-Pb data taking, the beam direction was reversed between runs 211256 and 211313. The list of validated runs, which must be applied to all analyses, either with the full validation or for an analysis requiring only `muons`, can be found in [Validated runs, full validation](#).
[Validated runs, muons only](#)

Related datasets
The corresponding PromptReco-v1 dataset: [/PAMuon/HIRun2013-PromptReco-v1/RECO](#)

Dataset characteristics
15445691 events, 1636 files, 5.0 TiB in total.

Dataset characteristics
15445691 events, 1636 files, 5.0 TiB in total.
20 files available. Request to access the rest

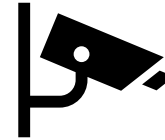
Work in progress:
Still in the design phase



Cold storage challenges

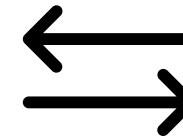
1. Monitoring

1. Identify candidates for cold storage
2. Keep track of cold storage requests



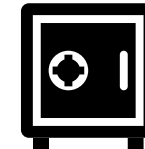
2. Transfer files

- A. CERN Tape Archive: FTS
- B. Experiment frameworks: Dedicated plugins per experiment



3. Data sovereignty

1. Experiment still responsible for data (!)



4. Multiple copies of datasets

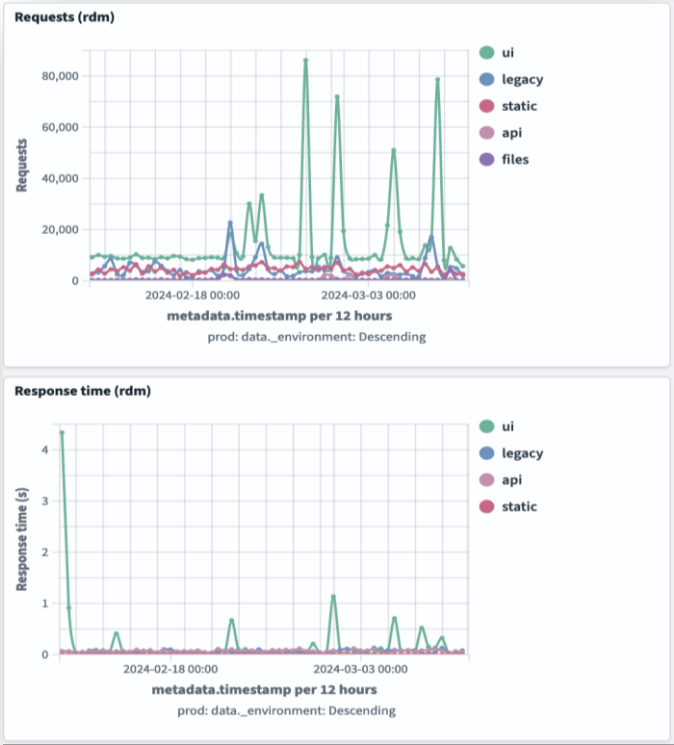
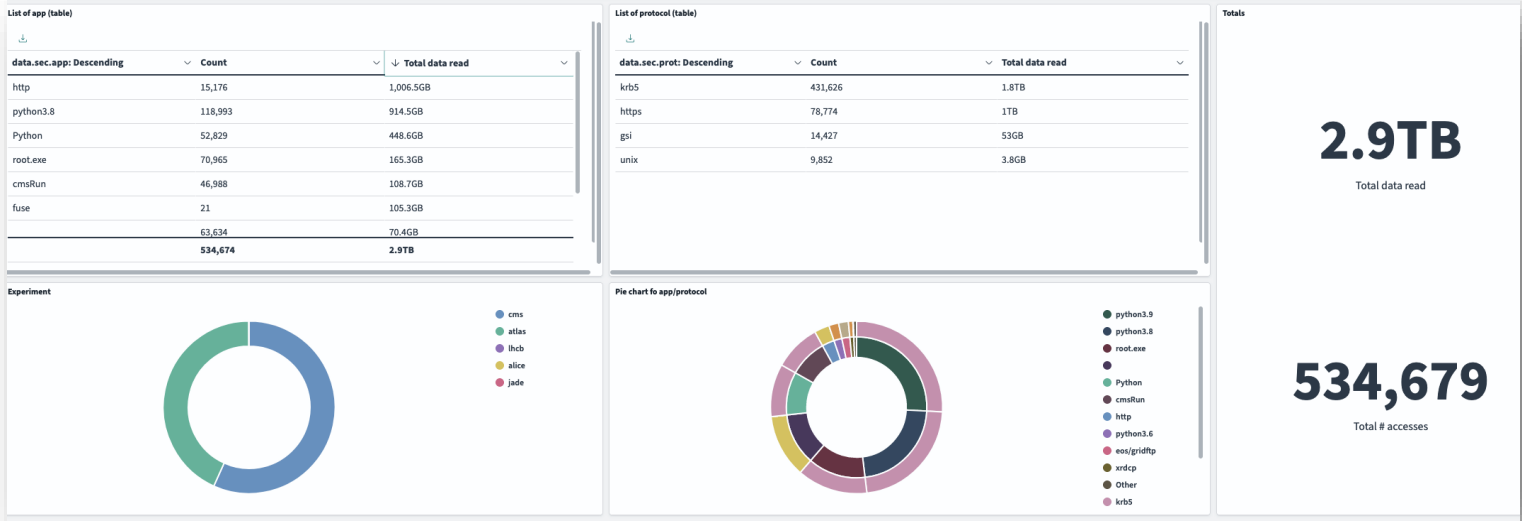


Monitoring

Framework provided by CERN MONIT

Part of the data sent by Storage group

Last 30 days available



CERN Open Data Portal

- Bringing experiment data to the general public
 - Policies with embargo period

- Digital repository with 4.5 PB of data
- Following FAIR guidelines



- Curated data: datasets, documentation, software...
- Evaluating cold storage integration

opendata-team@cern.ch

<http://opendata.cern>



home.cern