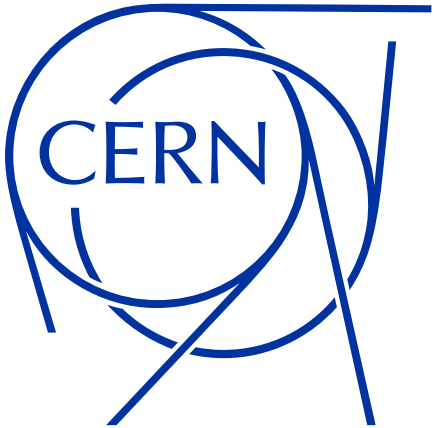


CERN Open Data

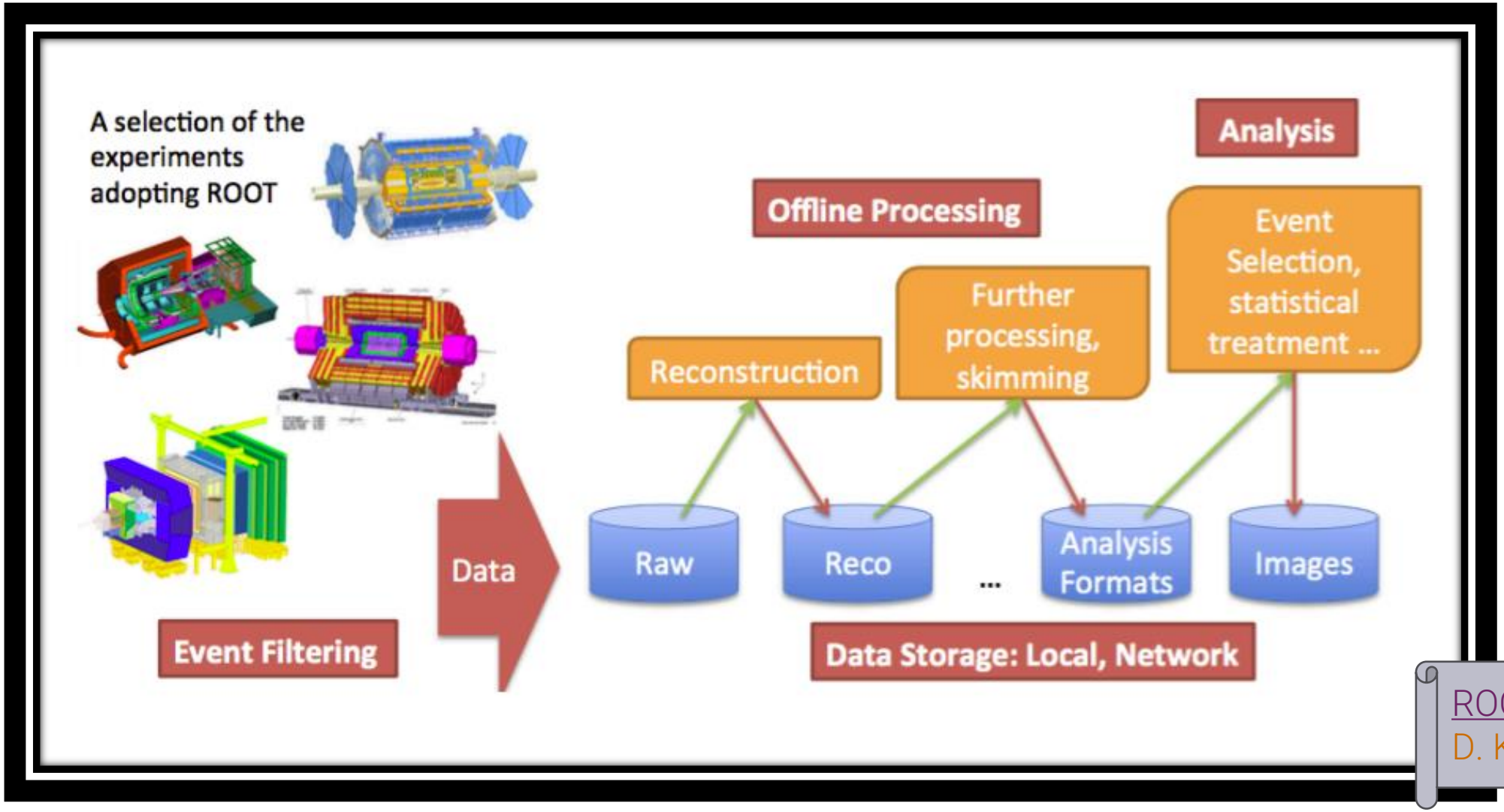
CS3 2024

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 lay the foundations 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 collaborators, the wider scientific community and the general public to analyse data for educational purpose and for eventual reassessment of the published results. The present document describes the basic principles that will guide the redaction 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 conditions parameters are needed to transform raw detector information into physics information, free of detector biases. The various ALICE data formats are classified as follows:

a) Raw data embedding the signals delivered by the detectors along with the associated status data containing various information on the running conditions constitutes the primary information collected by the ALICE experiment. They provide the input of the reconstruction algorithm, together with the calibration data stored in a dedicated database;
b) 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;
c) Event Summary Data (ESD) produced by the reconstruction algorithm, 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;
d) 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;
e) Custom analysis object data, used standalone or together with the general purpose AOD for specific analysis;
f) 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.


Approved: CB 20th June 2014

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 as regards open access to data at different levels as described in the DPHEP [1] model. The main objective is to make the data available in a usable 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. [1]. This policy pertains to collision physics data (i.e. that are stored offline 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 is 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 encapsulation of the analysis is often provided for measurements in the framework of RIVET [3]. For searches information on signal acceptances 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 simplified, portable and self-contained formats for


CMS data preservation, re-use and open access policy

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 level of abstraction and at different points in time[1]. This opportunity calls for our collective responsibility and poses unprecedented challenges as no data sample of this 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 realisation 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 actors 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 0.9, 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 analysis workflows, 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 existing third-party platforms (e.g. INSPIRE[2]), 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 of their dependencies.

[1] A. Holmes, P. Igo-Kemenes, S. Metlay, "First results from the PARSE.Insight project: HEP survey on data preservation, re-use and (open) access" http://arxiv.org/abs/0906.0681
[2] http://inspirehep.net


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 jointly 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 DPHEP 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 stored offline 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.

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

CERN Open Data Portal

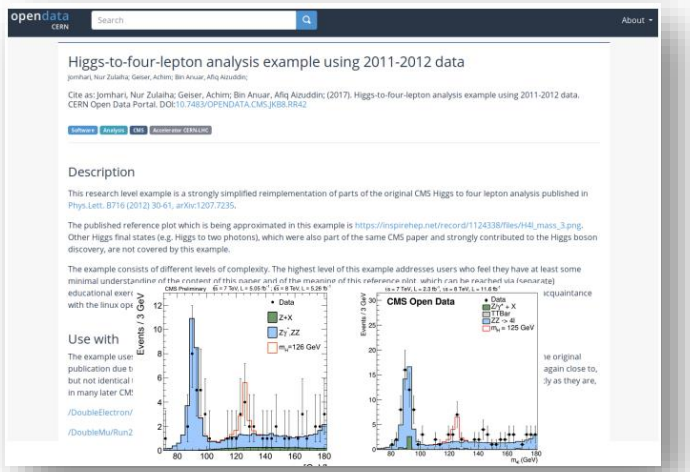
http://opendata.cern

- **Repository of data for the general public**
  - Launched in November 2014
  - Using Invenio Framework

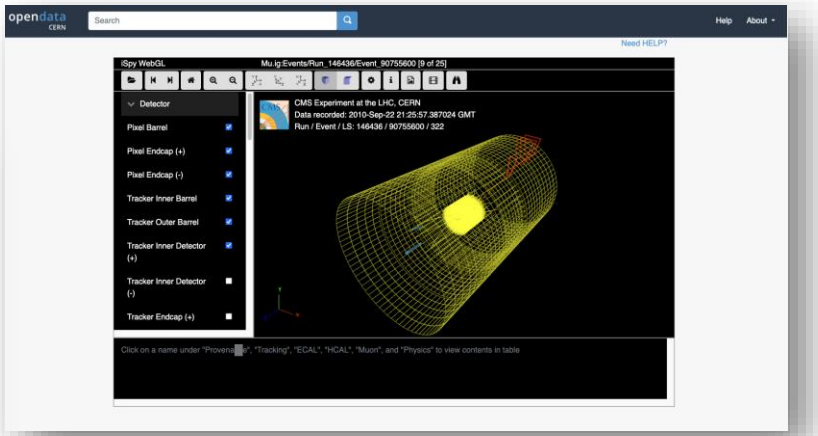
- **Plenty of content**
  - Dataset
    - Collision, simulated and derived
  - Documentation
    - Glossary, tutorials, configuration, examples
  - Software
    - Frameworks, virtual machines, containers
- **Current size (Mar 2024)**
  - > 24.000 records
  - > 1.900.000 files
  - > 4.5 PB

Developed by CERN in collaboration with Experiments

CERN Open Data

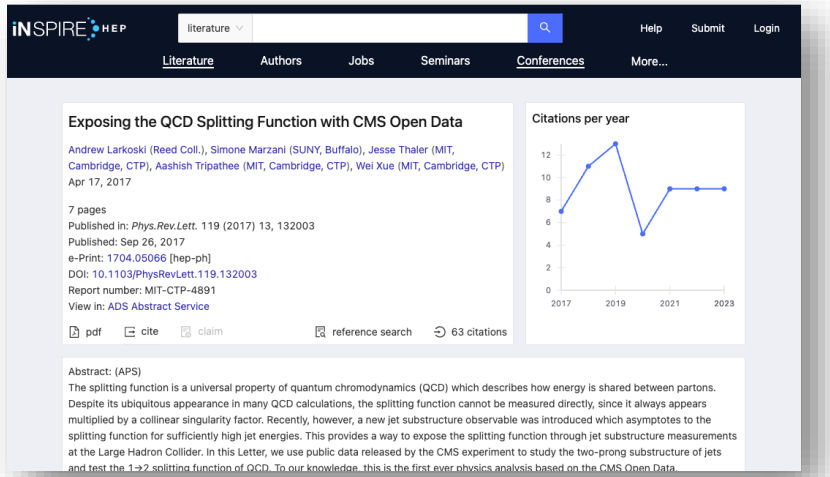


Run research-grade analysis examples



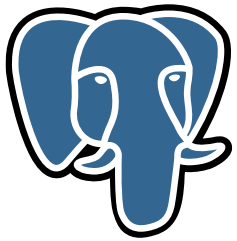
Interactive event display and histograms

Independent research…



… that the CMS collaboration cites

Technologies



FAIR guiding principles

- **F**indable
- **A**ccesible
- **I**nteroperable
- **R**eusable

Digital repositories for FAIR data management
Lars Holm Nielsen

zer10do
Celebrating our 10th Anniversary

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

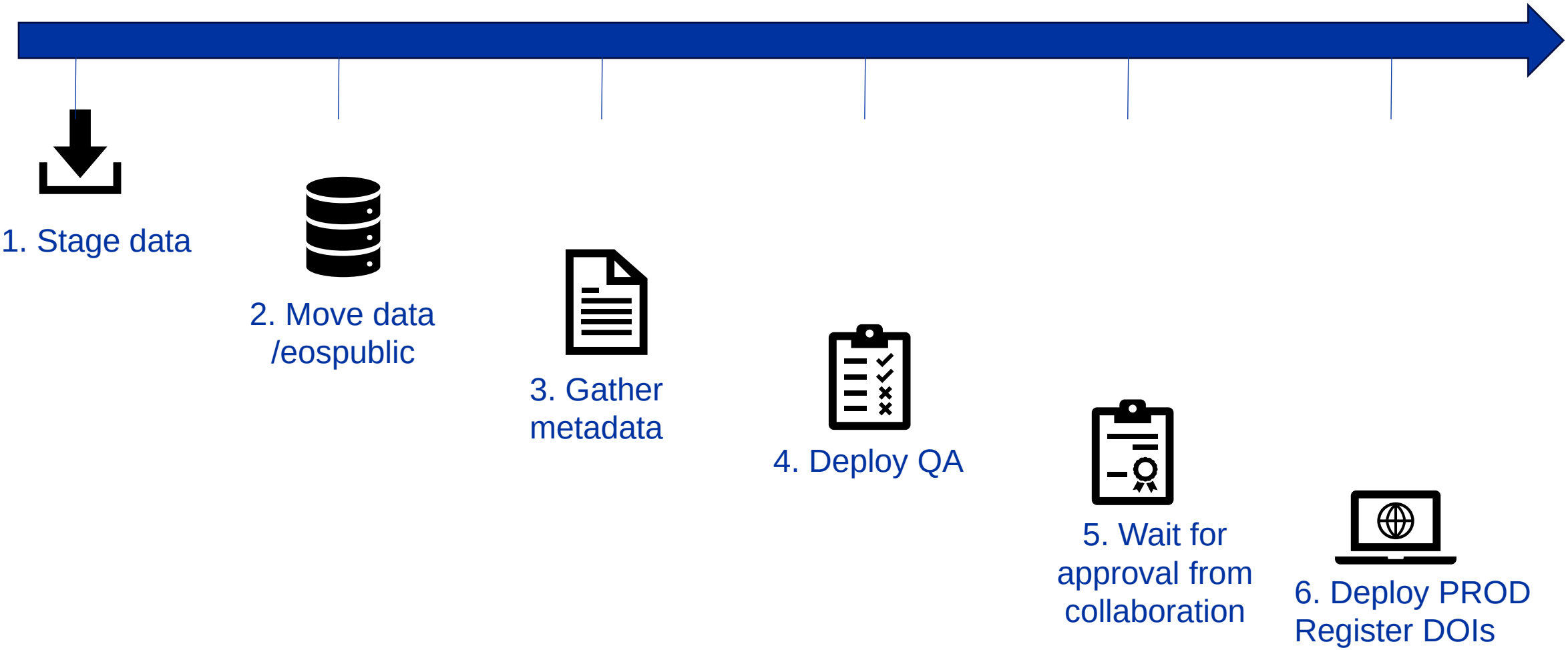
653k Accesses | **6374** Citations | **2138** Altmetric | Metrics

An Addendum to this article was published on 19 March 2019

Abstract

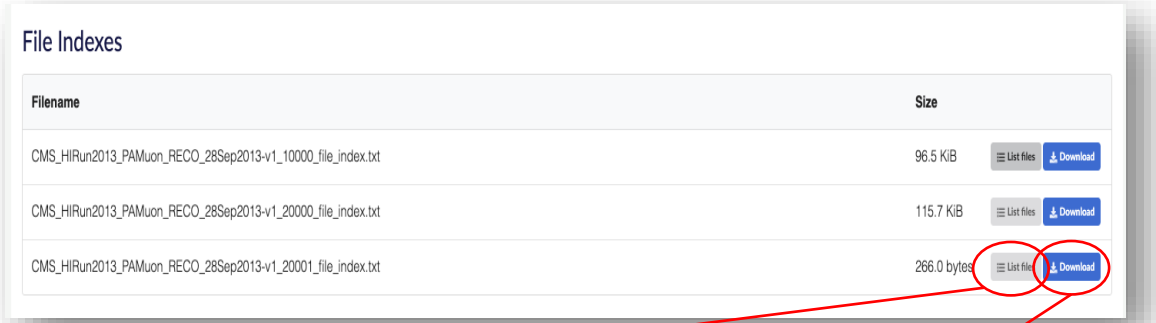
There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we 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.

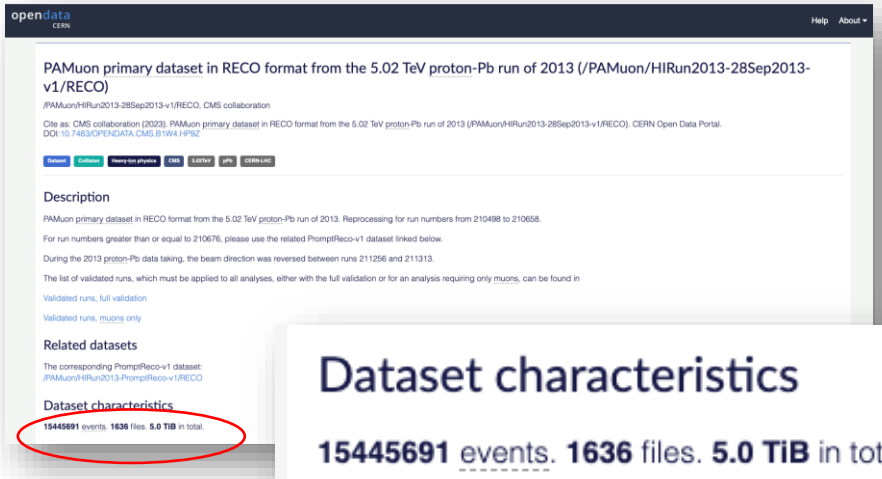
Curating content

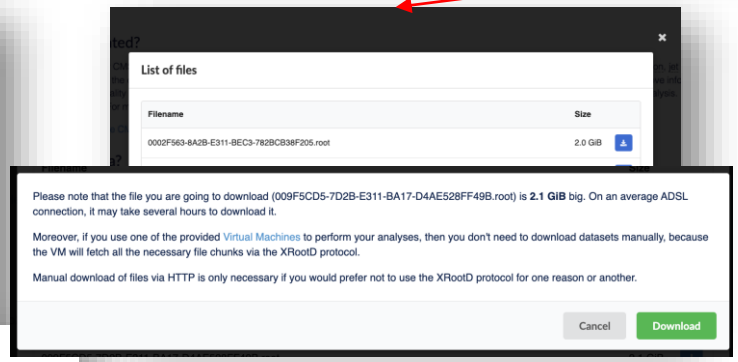


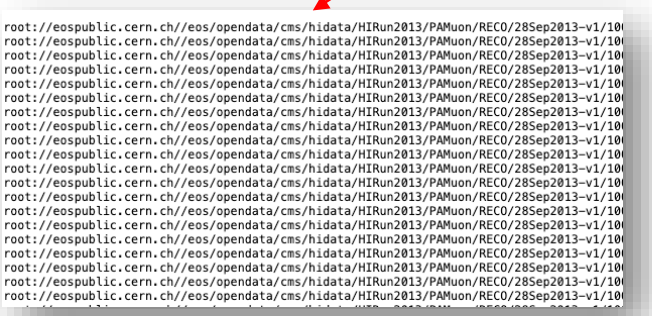
CERN Open Data dataset access

opendata CERN









Optimize storage cost

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

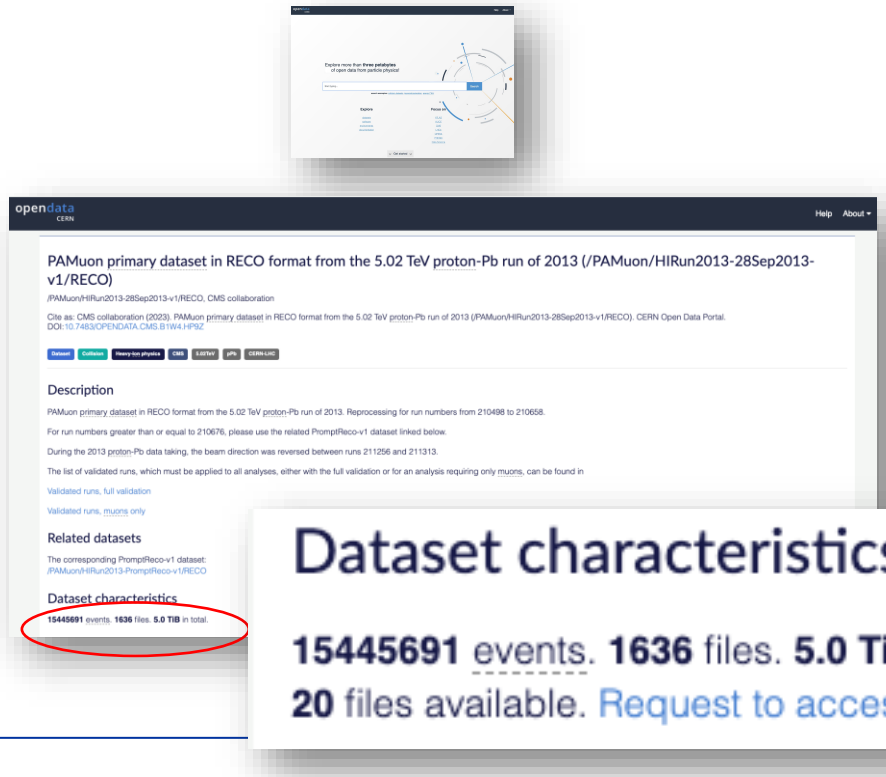
- **Are there any more efficient ways of long term archival?**

- **Possibilities:**
  - Computing instead of storage (LHCb Ntupling service)
  - Cold Storage. Move part of the data to cheaper storage (Tape, experiment framework)

Cold Storage integration

opendata CERN

Work in progress:
Still in the design phase



EOS

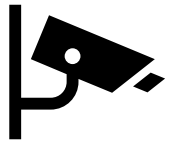
CERN Tape Archive

FTS File Transfer Service

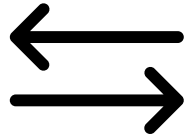
Experiment Data Framework

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 (!)

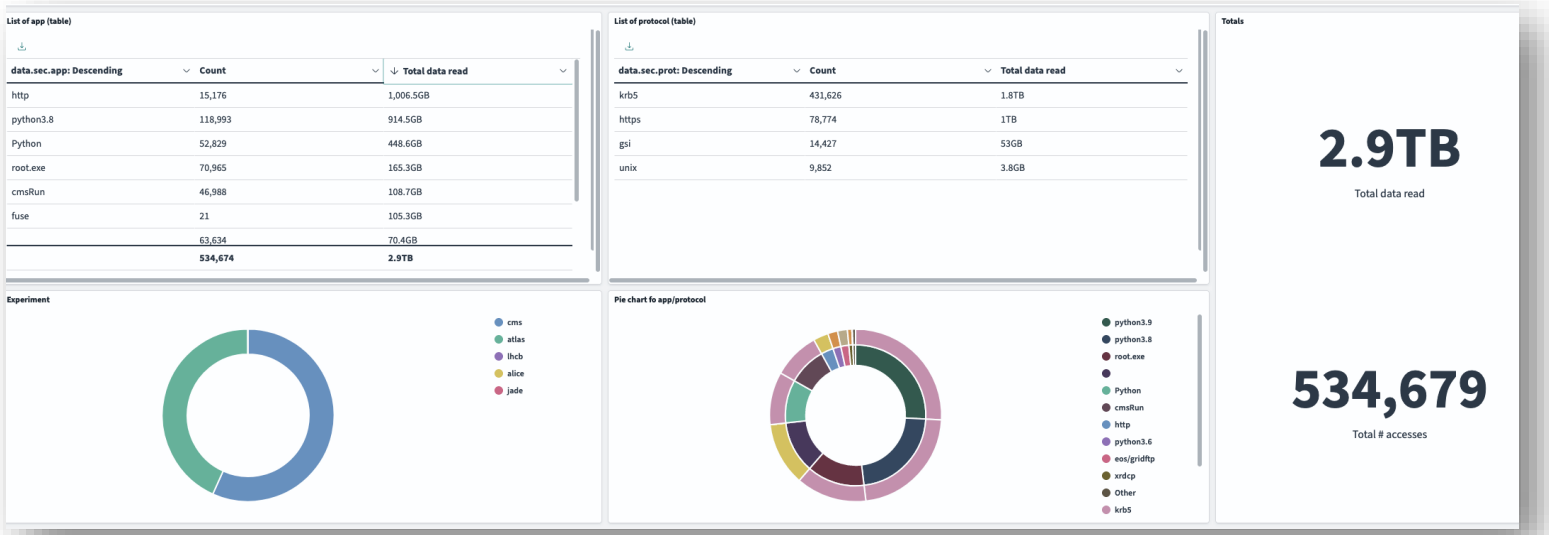
4. **Multiple copies of datasets**

Monitoring

Framework provided by CERN MONIT

Part of the data sent by Storage group

Last 30 days available



List of app (table)

data.sec.app: Descending	Count	Total data read
http	15,176	1,006.5GB
python3.8	118,993	914.5GB
Python	52,829	448.6GB
root.exe	70,965	165.3GB
cmsRun	46,988	108.7GB
fuse	21	105.3GB
	63,634	70.4GB
	534,674	**2.9TB**

List of protocol (table)

data.sec.prot: Descending	Count	Total data read
krb5	431,626	1.8TB
https	78,774	1TB
gsi	14,427	53GB
unix	9,852	3.8GB

Totals

2.9TB

Total data read

534,679

Total # accesses

Experiment

Pie chart fo app/protocol

Requests (rdm)

Response time (rdm)

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