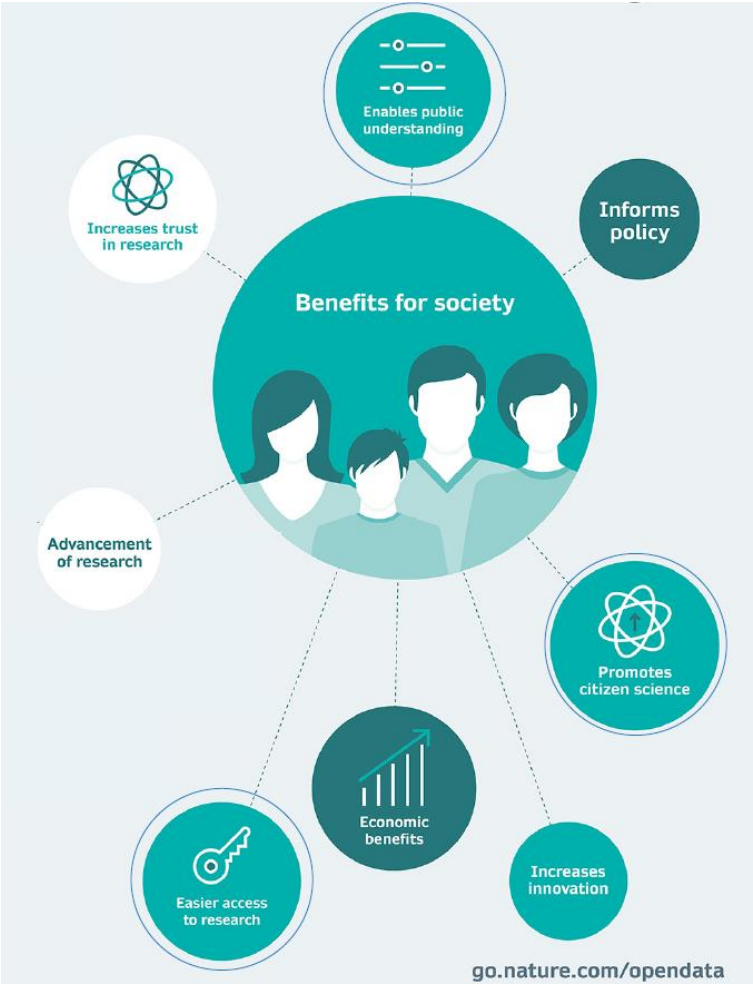
CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS

Instituto de Ciencias Nucleares UNAM

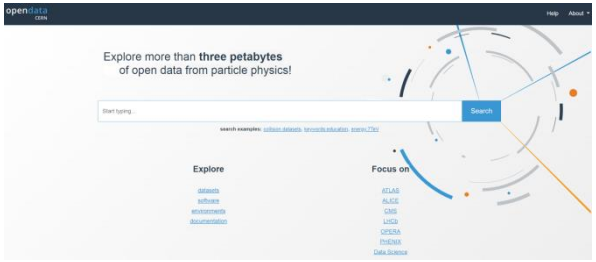
Fostering educational efforts in Latin America through the use of CERN Open Data

Dr. Leonid Serkin
Instituto de Ciencias Nucleares, UNAM, Mexico

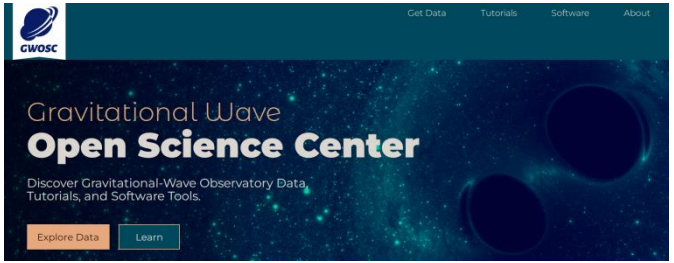


Open data is a crucial part of science and brings many benefits to society and the public:

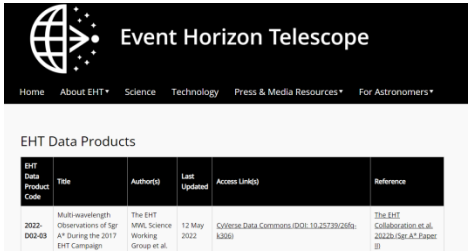
- Increases trust in scientific research.
- Allows the public to understand the results obtained.
- Promotes citizen science.
- Increases understanding of innovation and economic growth.
- Fosters educational efforts, allowing students and teachers to obtain and reproduce our results.



opendata.cern.ch



gwosc.org



https://eventhorizontelescope.org/for-astronomers/data

2012: Discovery of the Higgs boson at the LHC at CERN.

2015: Detection of gravitational waves by LIGO.

2019: Direct detection of the event horizon of a black hole by EHT.

Open Data is available for all of them!

nature physics

PERSPECTIVE

https://doi.org/10.1038/s41567-018-0342-2

Corrected: Publisher Correction

OPEN

Open is not enough

Xiaoli Chen[1,2], Sünje Dallmeier-Tiessen[1*], Robin Dasler[1,3], Sebastian Feger[1,3], Pamfilos Fokianos[1], Jose Benito Gonzalez[1], Harri Hirvonsalo[1,4,12], Dinos Kousidis[1], Artemis Lavasa[1], Salvatore Mele[1], Diego Rodriguez Rodriguez[1], Tibor Šimko[1*], Tim Smith[1], Ana Trisovic[1,5*], Anna Trzcinska[1], Ioannis Tsanaktsidis[1], Markus Zimmermann[1], Kyle Cranmer[6], Lukas Heinrich[6], Gordon Watts[7], Michael Hildreth[8], Lara Lloret Iglesias[9], Kati Lassila-Perini[4] and Sebastian Neubert[10]

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be embraced more widely. This first experience suggests that reproducibility requires going beyond openness.

Open science and reproducible research have become pervasive goals across research communities, political circles and funding bodies[1–3]. The understanding is that open and reproducible research practices enable scientific reuse, accelerating future projects and discoveries in any discipline. In the struggle to take concrete steps in pursuit of these aims there has been much discussion and awareness-raising, often accompanied by a push to make research products and scientific results open quickly.

Although these are laudable and necessary first steps, they are not sufficient to bring about the transformation that would allow us to reap the benefits of open and reproducible research. It is time to move beyond the rhetoric and the trust in quick fixes and start designing and implementing tools to power a more profound change.

Our own experience from opening up vast volumes of data is that openness cannot simply be tacked on as an afterthought at the end of the scientific endeavour. In addition, openness alone does not guarantee reproducibility or reusability, so it should not be pursued as a goal in itself. Focusing on data is also not enough: it needs to be accompanied by software, workflows and explanations, all of which need to be captured throughout the usual iterative and closed research lifecycle, ready for a timely open release with the results.

Thus, we argue that having the reuse of research results as a goal requires the adoption of new research practices during the data analysis process. Such practices need to be tailored to the needs of each given discipline with its particular research environment, culture and idiosyncrasies. Services and tools should be developed with the idea of meshing seamlessly with existing research procedures, encouraging the pursuit of reusability as a natural part of researchers' daily work (Fig. 1). In this way, the generated research products are more likely to be useful when shared openly.

In tackling the challenge of enabling reusable research, we keep these ideas as our guiding light when putting changes into practice in our community—high-energy physics (HEP). Here, we illustrate our approach, particularly through our work at CERN, and present our community's requirements and rationale. We hope that the explanation of our challenges and solutions will stimulate discussions around the practical implementation of workflows for reproducible and reusable research more widely in other scientific disciplines.

Approaching reproducibility and reuse in HEP

To set the stage for the rest of this piece, we first construct a more nuanced spectrum in which to place the various challenges facing HEP, allowing us to better frame our ambitions and solutions. We choose to build on the descriptions introduced by Carole Goble[2] and Lorena A. Barba[3] shown in Table 1.

These concepts assume a research environment in which multiple labs have the equipment necessary to duplicate an experiment, which essentially makes the experiments portable. In the particle physics context, however, the immense cost and complexity of the experimental set-up essentially make the independent and complete replication of HEP experiments unfeasible and unhelpful. HEP experiments are set up with unique capabilities, often being the only facility or instrument of their kind in the world; they are also constantly being upgraded to satisfy requirements for higher energy, precision and level of accuracy. The experiments at the Large Hadron Collider (LHC) are prominent examples. It is this uniqueness that makes the experimental data valuable for preservation so that it can be later reused with other measurements for comparison, confirmation or inspiration.

Our considerations here really begin after gathering the data. This means that we are more concerned with repeating or verifying the computational analysis performed over a given dataset rather than with data collection. Therefore, in Table 2 we present a variation of these definitions that takes into account a research environment in which 'experimental set-up' refers to the implementation of a computational analysis of a defined dataset, and a 'lab' can be thought of as an experimental collaboration or an analysis group.

In the case of computational processes, physics analyses themselves are intrinsically complex due to the large data volume and algorithms involved[4]. In addition, the analysts typically study more than one physics process and consider data collected under different running conditions. Although comprehensive documentation on the analysis methods is maintained, the complexity of the software implementations often hides minute but crucial details, potentially leading to a loss of knowledge concerning how the results were obtained[5].

In absence of solutions for analysis capture and preservation, knowledge of specific methods and how they are applied to a given physics analysis might be lost. To tackle these community-specific challenges, a collaborative effort (coordinated by CERN, but involving the wider community) has emerged, initiating various projects, some of which are described below.

Reuse and openness. The HEP experimental collaborations operate independently of each other, and they do not share physics results until they have been rigorously verified by internal review processes[6]. Because these reviews often involve the input of the entire collaboration, where the level of crosschecking is extensive, the measurements are considered trustworthy.

[1]CERN, Geneva, Switzerland. [2]Sheffield University, Sheffield, UK. [3]Stuttgart University, Stuttgart, Germany. [4]Helsinki Institute of Physics, Helsinki, Finland. [5]Cambridge University, Cambridge, UK. [6]NYU, New York, NY, USA. [7]University of Washington, Seattle, WA, USA. [8]University of Notre Dame, Notre Dame, IN, USA. [9]Instituto de Física de Cantabria CSIC-UC, Santander, Spain. [10]Heidelberg University, Heidelberg, Germany. [12]Present address: DataCite, German National Library of Science and Technology, Hanover, Germany. *e-mail: sunje.dallmeier-tiessen@cern.ch; tibor.simko@cern.ch; ana.trisovic@cern.ch

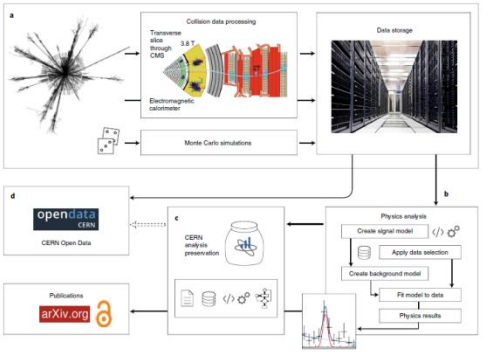


Fig. 1 | Data continuum in LHC experiments. a, The experimental data from proton–proton collisions in the Large Hadron Collider are being collected by particle detectors run by the experimental collaborations ALICE, ATLAS, CMS and LHCb. The raw experimental data is further filtered and processed to give the collision dataset formats that are suitable for physics analyses. In parallel, the computer simulations are being run in order to provide necessary comparison of experimental data with theoretical predictions. **b**, The stored collision and simulated data are then released for individual physics analyses. A physicist may perform further data reduction and selection procedures, which are followed by a statistical analysis on the data. Physics results are derived taking into account statistical and systematic uncertainties. The results often summarize which theoretical models have predictions that are consistent with the observations once background estimates have been included. The analysis assets being used by the individual researcher include the information about the collision and simulated datasets, the detector conditions, the analysis code, the computational environments, and the computational workflow steps used by the researcher to derive the histograms and the final plots as they appear in publications. **c**, The CERN Analysis Preservation service captures all the analysis assets and related documentation via a set of 'push' and 'pull' protocols, so that the analysis knowledge and data are preserved in a trusted long-term digital repository for preservation purposes. **d**, The CERN Open Data service publishes selected data as they are released by the LHC collaborations into the public domain after an embargo period of several years depending on the collaboration data management plans and preservation policies. Credit: CERN (**a**); Dawe Gandy (**b**,**c**, code icon); Simpleicon (**b**,**c**, gear icon); Andrian Valeanu (**b**,**c**, data icon); Umar Irshad (**c**, paper icon); Freepik (**c**, workflow icon).

Table 1 | Terminology related to reproducible research introduced by Carole Goble and Lorena A. Barba

Term	Purpose	Description
Rerun	Robust	Variations on experiment and set-up, conducted in the same lab
Repeat	Defend	Same experiment, same set-up, same lab
Replicate	Certify	Same experiment, same set-up, independent lab
Reproduce	Compare	Variations on experiment and set-up, independent labs
Reuse	Transfer	Different experiment

Nature Physics **15**, 113–119 (2019)

Open is not enough:

- define your reproducibility.
- runnable usage examples.
- preserve computational workflow.
- enable FAIR reuse

scientific **data**

Explore content ▾ About the journal ▾ Publish with us ▾

nature › scientific data › comment › article

Comment | 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

732k Accesses | **5469** Citations | **2223** 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

Scientific Data 3, 160018 (2016)

Box 2 | The FAIR Guiding Principles

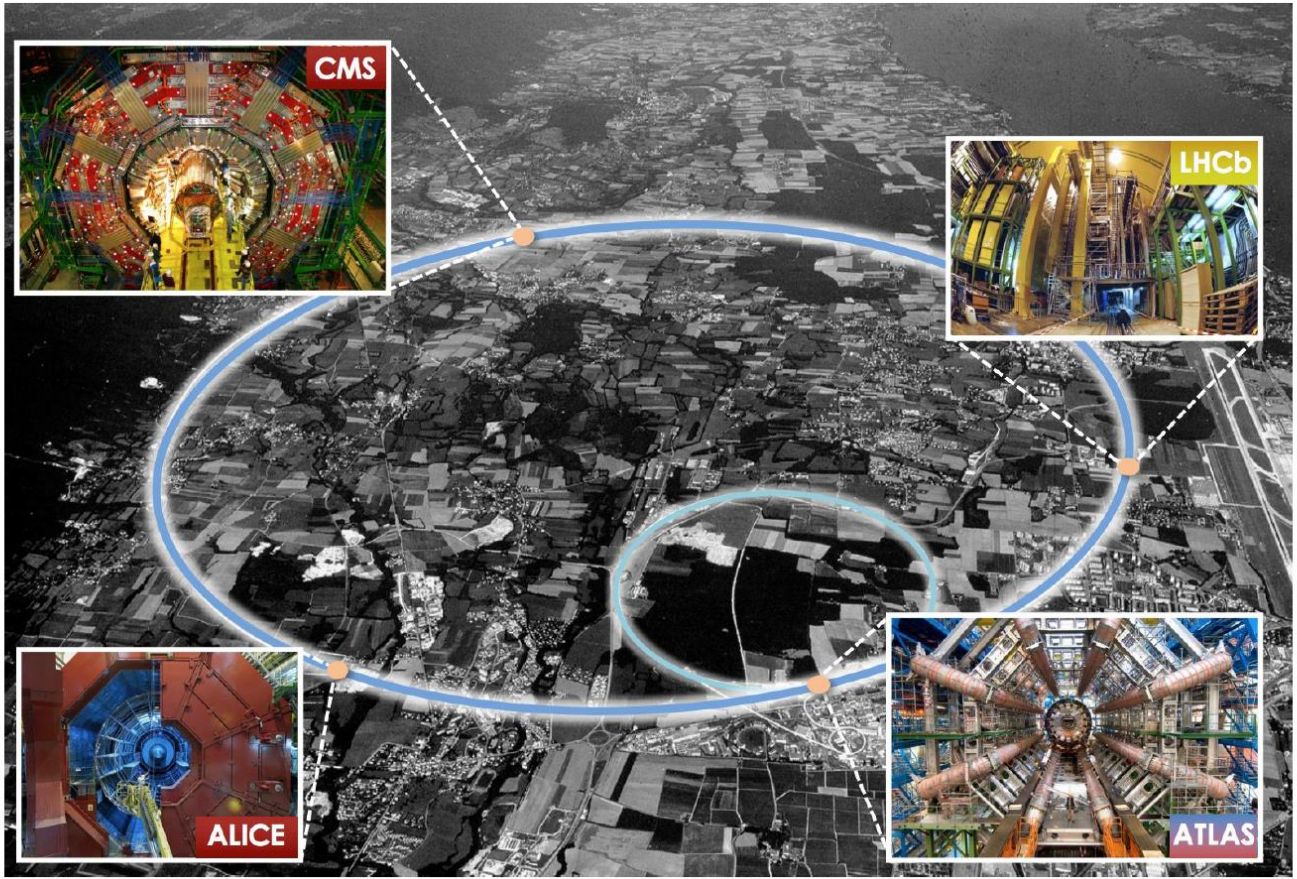
To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Brazil
Colombia
Ecuador
Mexico



Brazil
Colombia

Brazil
Cuba
Mexico
Peru

Argentina
Brazil
Chile
Colombia

LHC data collection:

• 90 petabytes (PB = 1024 TB) per year from LHC + 25 PB from other experiments

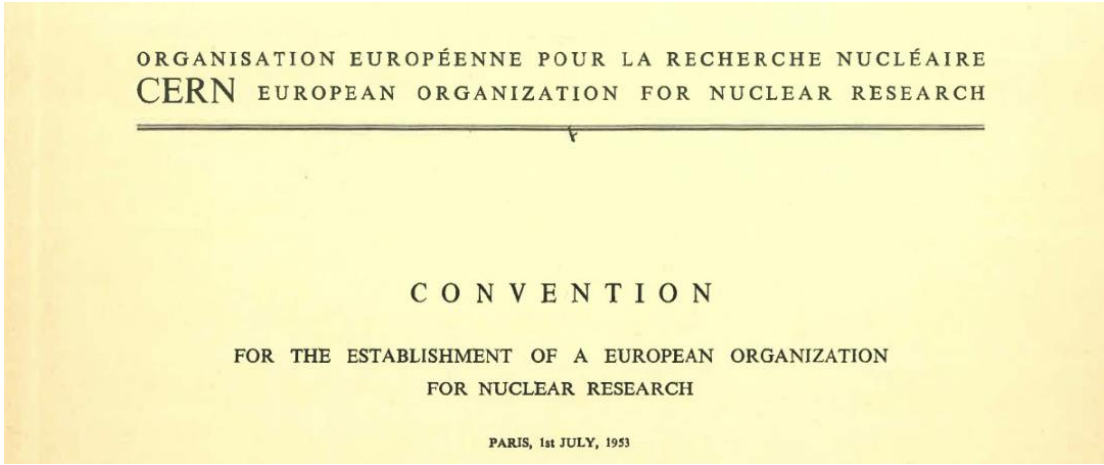
LHC data storage:

• CERN data center has ~400 PB on tapes + ~350 PB on disks

• CERN storage system served 2.5 exabyte (EB = 1024 PB) of data (just in 2020!)

"…and the results of its experimental and theoretical work shall be published or otherwise made generally available"

CERN Founding Convention (1953)



ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CONVENTION

FOR THE ESTABLISHMENT OF A EUROPEAN ORGANIZATION
FOR NUCLEAR RESEARCH

PARIS, 1st JULY, 1953

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 embod unique scientific information for the in-depth understanding of the profound nature and origin of matter. Because of their unique-ness, long term preservation must be an essential objective of the data processing framework and will lay the foun-dations of the ALICE Collaboration legacy to the scientific community as well as to the general public. These consid-erations 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 analyze data for educational purpose and for eventual reassessment of the published results. The present document describes the basic principles that will guide the notation addressed by the ALICE data preservation policy.

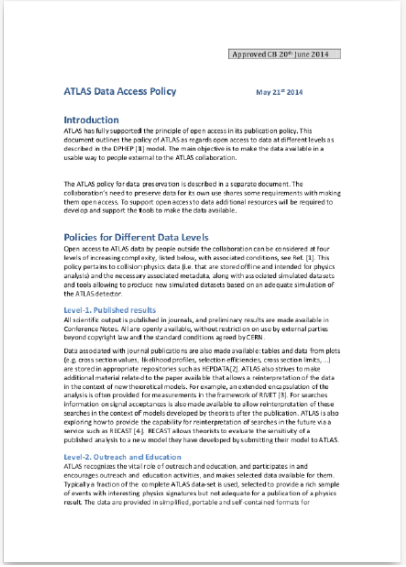
ALICE data formats

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 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 algorithms, for both Monte Carlo and raw data. The ESD events provide calibrated tracks in a generic format, but also additional detector specific infor-mation allowing a full physics analysis;
d) General purpose *Analysis Object Data* (AOD), derived from ESD data. The AOD data format contains a sim-plified 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 analy-sis;
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 [3] 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. [3]. 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 recognises 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

Version 1.2
Approved by the CMS Collaboration Board 30th April 2018

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.

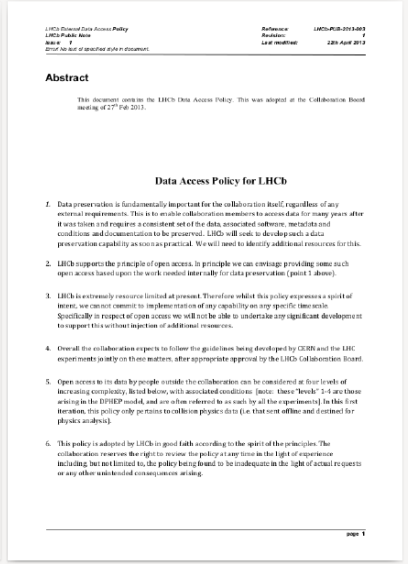
CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. To that extent, CMS will provide open access to its data after a suitable but relatively short embargo period, allowing CMS collaborators to fully exploit their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

This policy describes the CMS principles of data preservation, re-use and open access, as well as the relevant 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), 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. Metz, "Post results from the PARSE.Insight project: HEP survey on data preservation, re-use and (open) access" http://arxiv.org/abs/0906.0485
[2] http://inspirehep.net



LHCb-Public Note — Reference: LHCb-PUB-2013-003
Issue: 1 — Revision: 1
Error! No text of specified style in document. — Last modified: 23th April 2013

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 to require 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 sent 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.

page 1

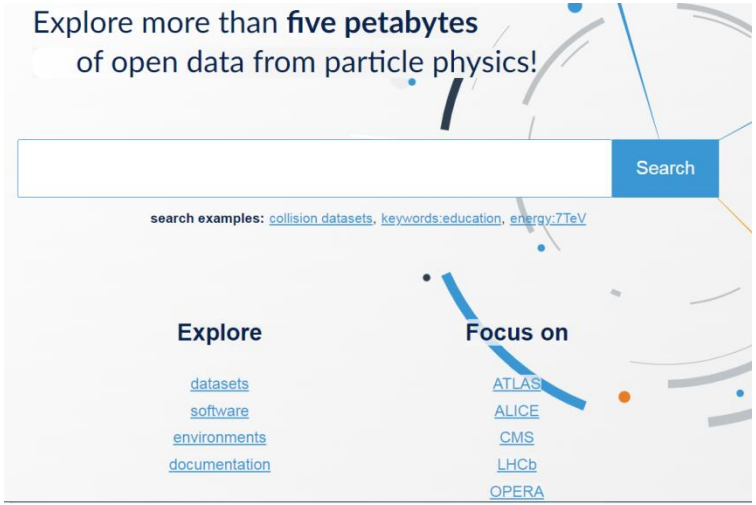
Restricted data → embargo period → open data

opendata CERN

https://opendata.cern.ch/

Launched in November 2014:

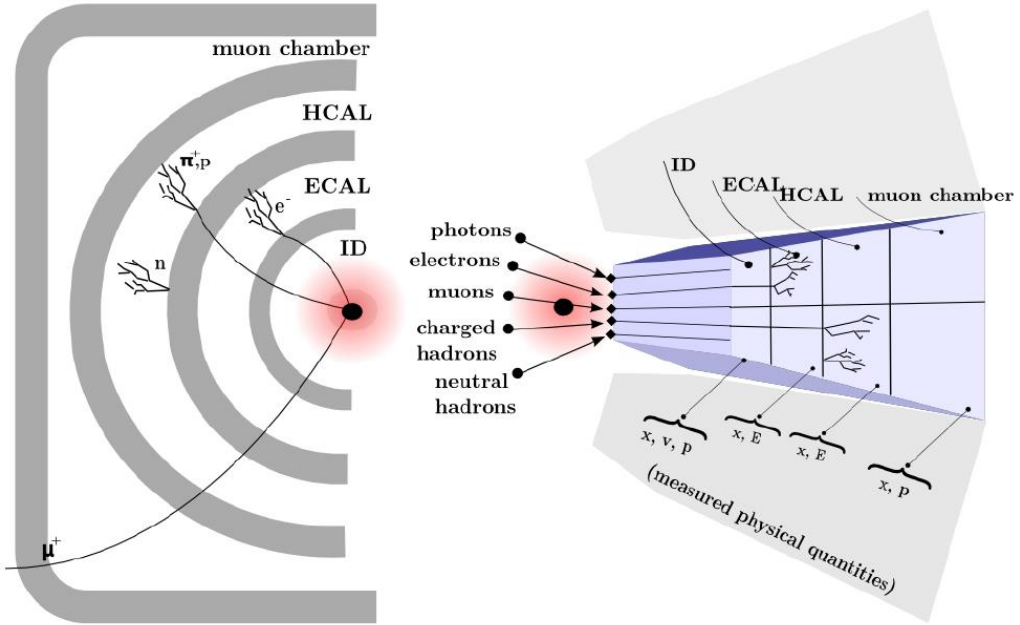
- Collision and simulated datasets for research
- Derived datasets for education
- Configuration files and documentation
- Virtual machines and container images
- Software tools and analysis examples

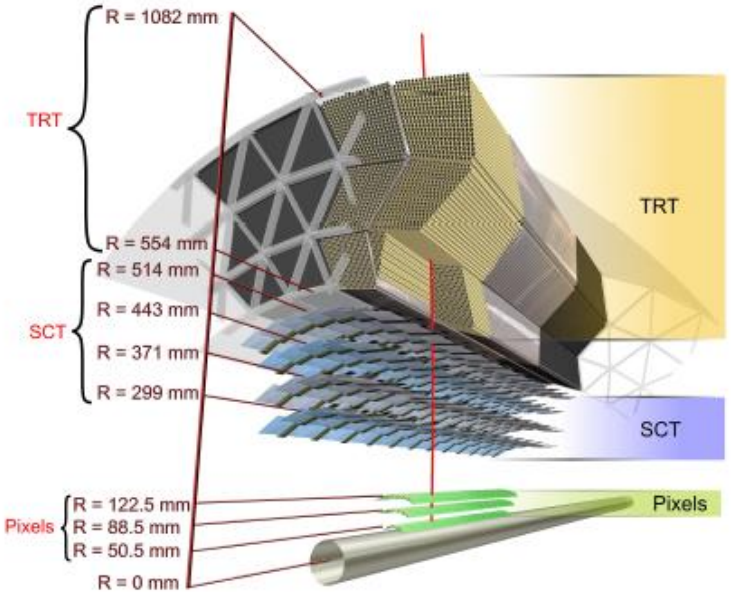


As of today:

- over 40.000 datasets
- over 50 different software
- over 5 petabytes of data

Each layer of the ATLAS detector is sensitive to different types of particles and/or radiation produced in the collision: photons, electrons, muons, charged and neutral hadrons...





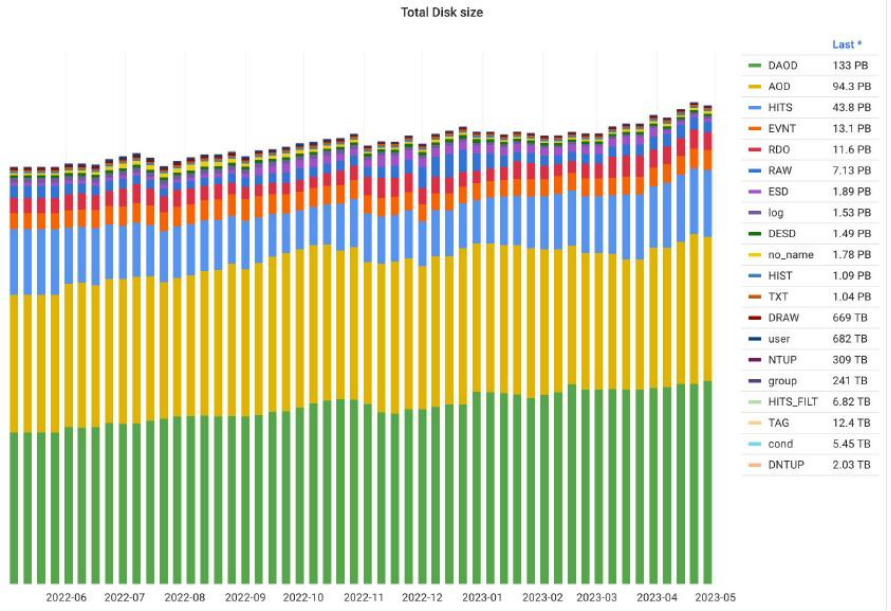
Data analyses require hundreds of petabytes of data storage and hundreds of thousands of CPUs to execute millions of lines of code.

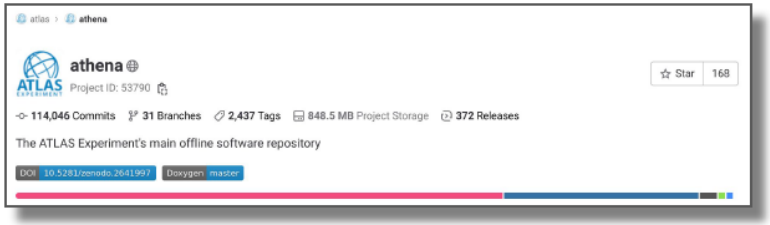
ATLAS has collected ~4 PB of data so far.

Data reconstruction and preprocessing take ~100 PB of space.

The software for reconstruction, simulation, and analysis consists of ~4 million lines of C++ code.

None of this is trivial to communicate and share with the public...





April 2024: CMS publishes proton collision data at 13 TeV from 2016. More than 70 TB of 13 TeV data and 830 TB of MC.

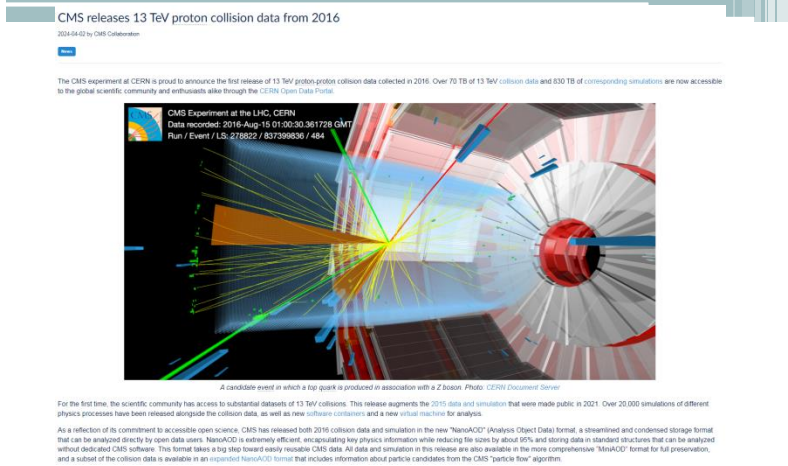
February 2024: The ATLAS Experiment releases 25% of the main proton-proton datasets for research purposes.

December 2023: LHCb releases the complete Run I dataset. 800 TB of data and algorithms for research and education.

December 2022: CMS completes the release of all its Run I proton-proton collision data. 491 TB of data and code.

February 2020: The ATLAS Experiment releases open data at 13 TeV for education.

November 2014: ALICE releases educational datasets. ALICE's datasets customized for demonstration and education.



CMS releases 13 TeV proton collision data from 2016

2024-04-02 by CMS Collaboration

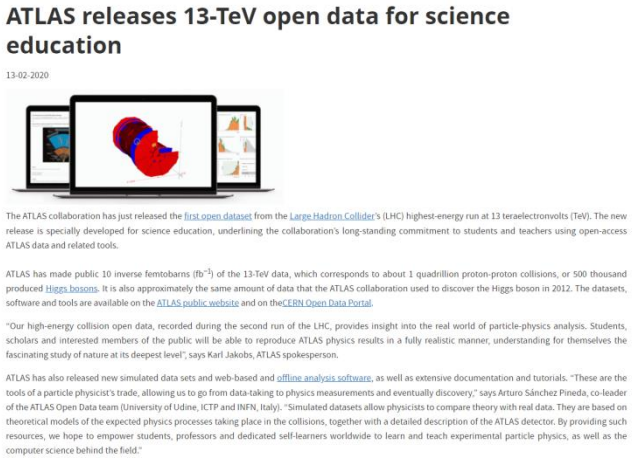
News

The CMS experiment at CERN is proud to announce the first release of 13 TeV proton-proton collision data collected in 2016. Over 70 TB of 13 TeV collision data and 830 TB of corresponding simulations are now accessible to the global scientific community and enthusiasts alike through the CERN Open Data Portal.

A candidate event in which a top quark is produced in association with a Z boson. Photo: CERN Document Server

For the first time, the scientific community has access to substantial datasets of 13 TeV collisions. This release augments the 2015 data and simulation that were made public in 2021. Over 20,000 simulations of different physics processes have been released alongside the collision data, as well as new software containers and a new virtual machine for analysis.

As a reflection of its commitment to accessible open science, CMS has released both 2016 collision data and simulation in the new "NanoAOD" (Analysis Object Data) format, a streamlined and condensed storage format that can be analyzed directly by open data users. NanoAOD is extremely efficient, encapsulating key physics information while reducing file sizes by about 95% and storing data in standard structures that can be analyzed without dedicated CMS software. This format takes a big step toward easily reusable CMS data. All data and simulation in this release are also available in the more comprehensive "MiniAOD" format for full preservation, and a subset of the collision data is available in an expanded NanoAOD format that includes information about particle candidates from the CMS "particle flow" algorithm.



ATLAS releases 13-TeV open data for science education

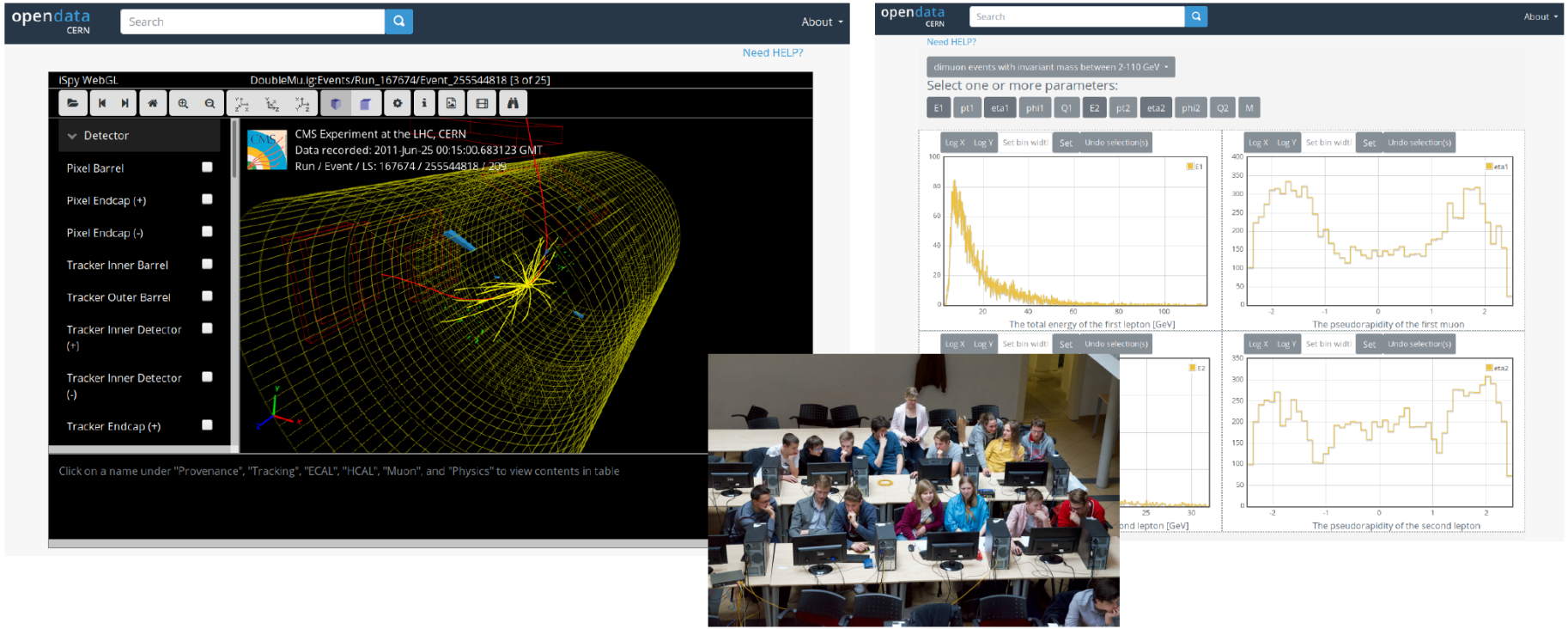
13-02-2020

The ATLAS collaboration has just released the first open dataset from the Large Hadron Collider's (LHC) highest-energy run at 13 teraelectronvolts (TeV). The new release is specially developed for science education, underlining the collaboration's long-standing commitment to students and teachers using open-access ATLAS data and related tools.

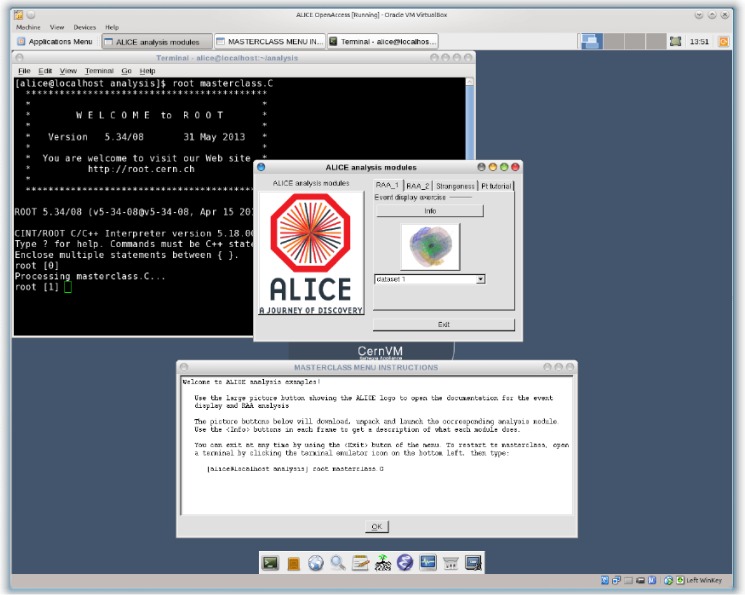
ATLAS has made public 10 inverse femtobarns (fb^{-1}) of the 13-TeV data, which corresponds to about 1 quadrillion proton-proton collisions, or 500 thousand produced Higgs bosons. It is also approximately the same amount of data that the ATLAS collaboration used to discover the Higgs boson in 2012. The datasets, software and tools are available on the ATLAS public website and on theCERN Open Data Portal.

"Our high-energy collision open data, recorded during the second run of the LHC, provides insight into the real world of particle-physics analysis. Students, scholars and interested members of the public will be able to reproduce ATLAS physics results in a fully realistic manner, understanding for themselves the fascinating study of nature at its deepest level", says Karl Jakobs, ATLAS spokesperson.

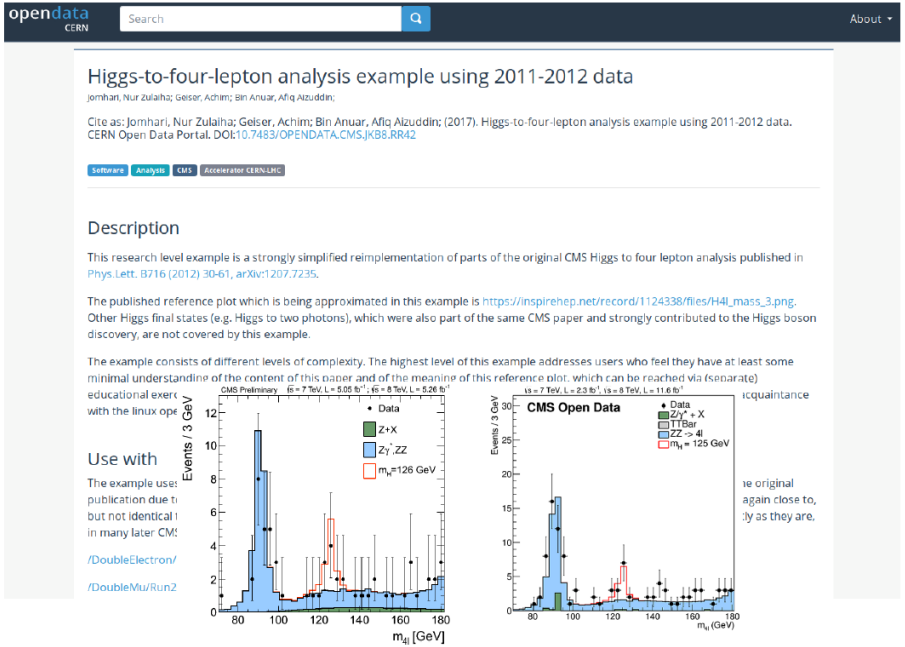
ATLAS has also released new simulated data sets and web-based and offline analysis software, as well as extensive documentation and tutorials. "These are the tools of a particle physicist's trade, allowing us to go from data-taking to physics measurements and eventually discovery," says Arturo Sánchez Pineda, co-leader of the ATLAS Open Data team (University of Udine, ICTP and INFN, Italy). "Simulated datasets allow physicists to compare theory with real data. They are based on theoretical models of the expected physics processes taking place in the collisions, together with a detailed description of the ATLAS detector. By providing such resources, we hope to empower students, professors and dedicated self-learners worldwide to learn and teach experimental particle physics, as well as the computer science behind the field."



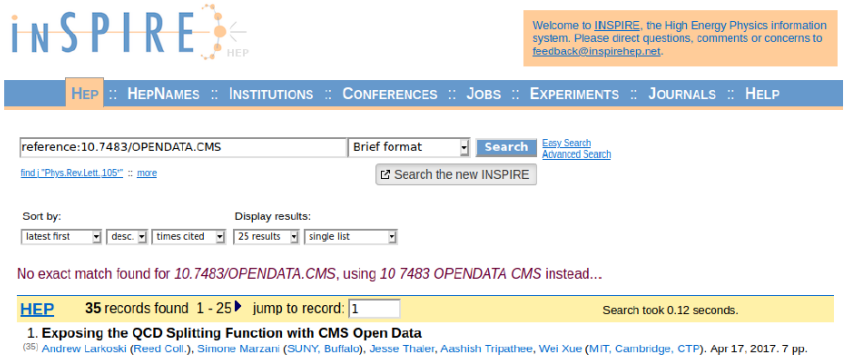
Interactive event display and histogram creation using small datasets

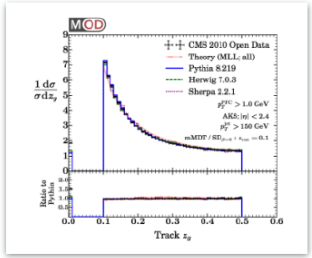


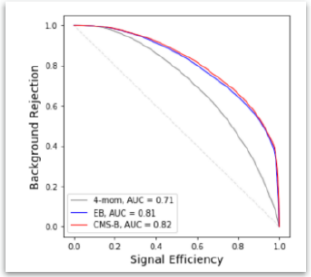
CERN Virtual Machines

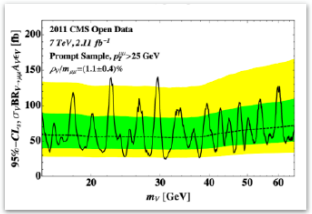


Simplified research-level analysis









arXiv:1704.05066 arXiv:1807.11916 arXiv:1902.04222

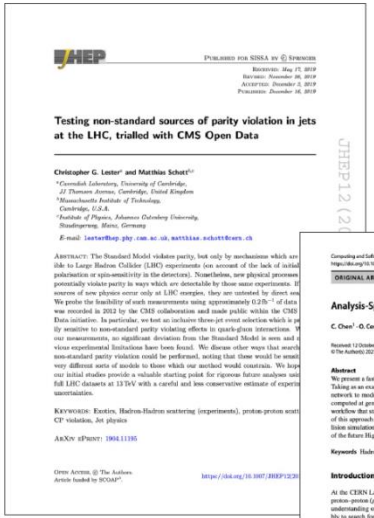
Searches, QCD jet studies, Machine Learning. . .

A measurement of the z_g distribution in pp collisions, using CMS open data, was recently reported [33, 34]. In PbPb collisions, this measurement reflects how the two color-charged par-

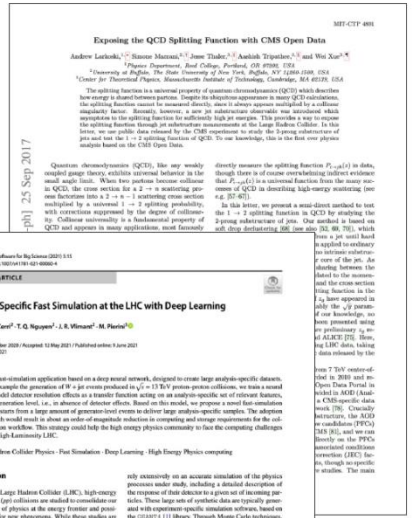
[33] A. Larkoski et al., "Exposing the QCD splitting function with CMS Open Data", *Phys. Rev. Lett.* **119** (2017) 132003, doi:10.1103/PhysRevLett.119.132003, arXiv:1704.05066.

[34] A. Tripathee et al., "Jet Substructure Studies with CMS Open Data", *Phys. Rev. D* **96** (2017) 074003, doi:10.1103/PhysRevD.96.074003, arXiv:1704.05842.

New theoretical ideas… being already cited by LHC experiments

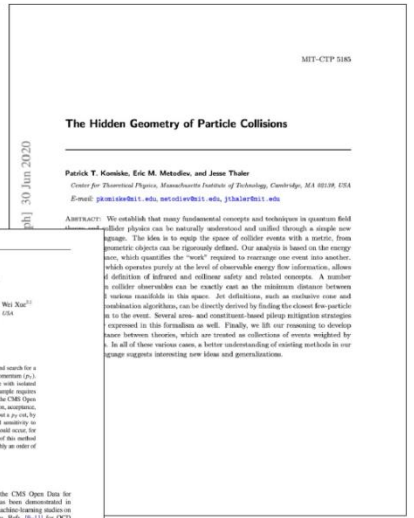


Testing non-standard sources of parity violation in jets at the LHC, trialled with CMS Open Data

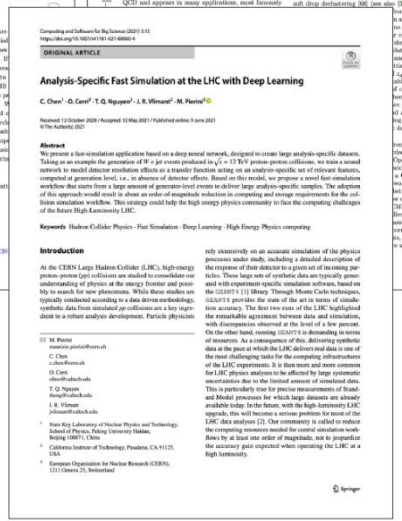


Exposing the QCD Splitting Function with CMS Open Data

Exploring the Space of Jets with CMS Open Data



The Hidden Geometry of Particle Collisions



Analysis-Specific Fast Simulation at the LHC with Deep Learning

Searching in CMS open data for dimson resonances with substantial transverse momentum

and much more

How can we overcome distances and enable anyone interested in particle physics and interaction of radiation with matter to learn remotely?

How can students replicate the procedures used at CERN?

How can teachers improve their own lectures and science communication?

The ATLAS Collaboration at CERN launched an educational platform (opendata.atlas.cern) designed to help students and teachers foster educational efforts at both undergraduate and graduate levels.



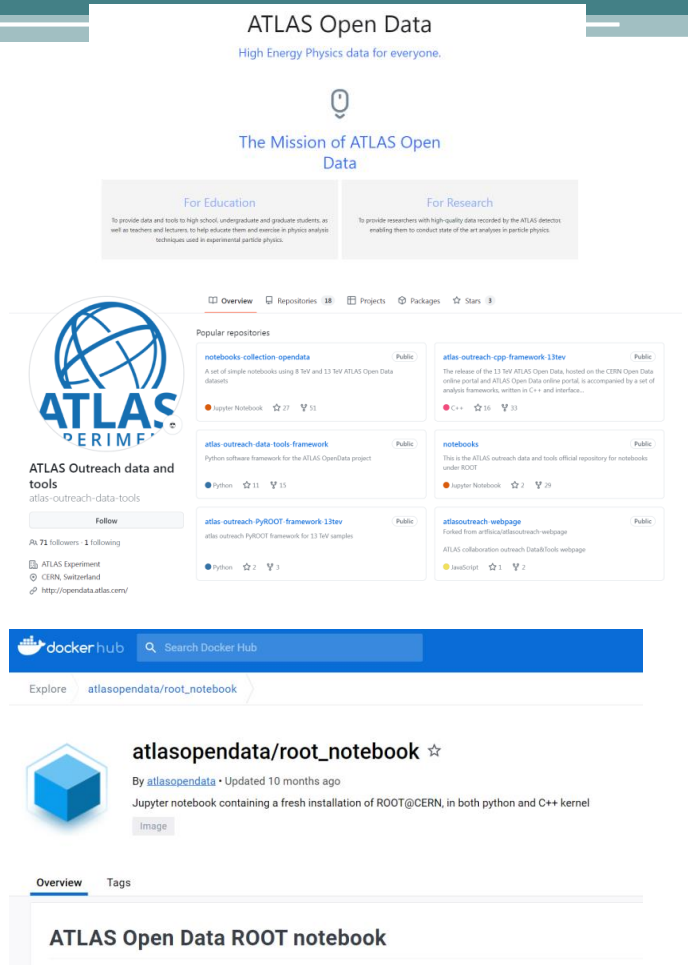


Open data stored at the CERN portal and on the online webpage

Comes with codes written in C++ and Python, uproot, pandas/numpy, pyROOT and RDataFrame, available on GitHub.

Interactive visual data analysis and notebooks Jupyter that allow one to get results online

Virtual machines that allow contains the operating system, software and data, and containers Docker for a local interface.



Recreate the main LHC discoveries over the past 10 years of data-taking:
find the heaviest top quark, or the Higgs boson in various final states

The audience

Advanced high school | Bachelors | Masters | PhD

~ hours | ~ days | ~ weeks | ~ semesters

Minimal setup & supervision | Hands-on setup & supervision

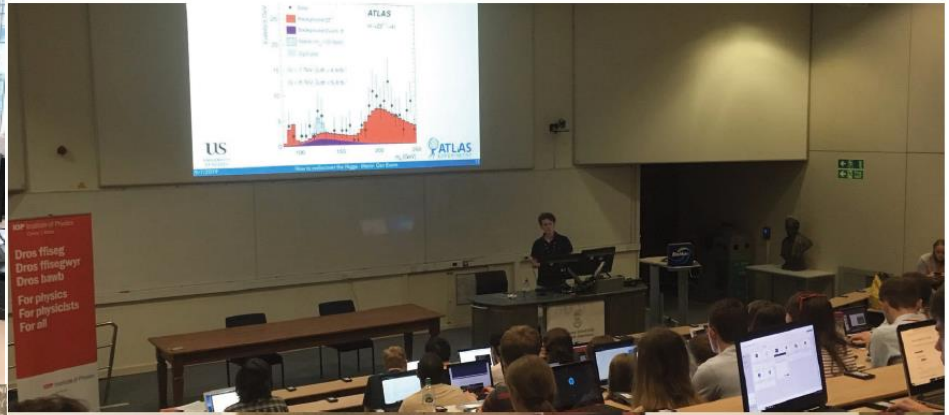
Browser based | Downloads required

CERN Open Data incorporated into the curriculum of several universities.

Educational workshops and schools using ATLAS Open Data worldwide, including Argentina, Colombia, Ecuador, Mexico, Peru, Uruguay, and Venezuela.

Educational workshops and schools using ATLAS Open Data worldwide, including Argentina, Colombia, Ecuador, Mexico, Peru, Uruguay, and Venezuela.

Educational workshops and schools using ATLAS Open Data worldwide, including Argentina, Colombia, Ecuador, Mexico, Peru, Uruguay, and Venezuela.

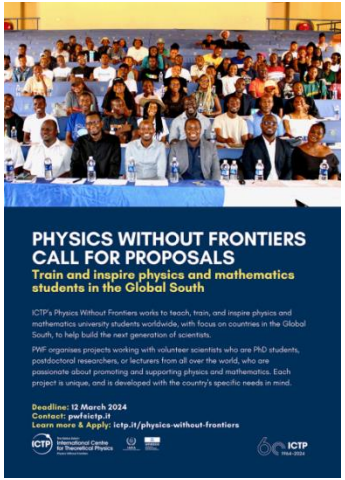


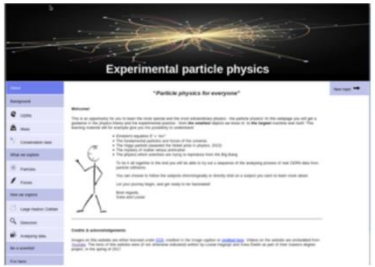
CERN International Teachers Programme

Open Data Challenge

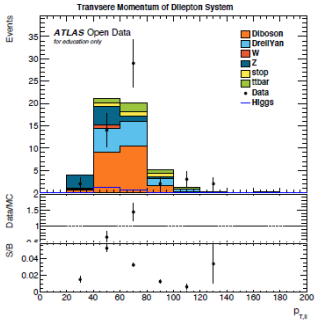
Science education programs in Latin America:

LA-CoNGA
ICTP Physics Without Frontiers
CEVALE2VE (Centro de Altos Estudios de Altas Energías)
PPGCosmo
…

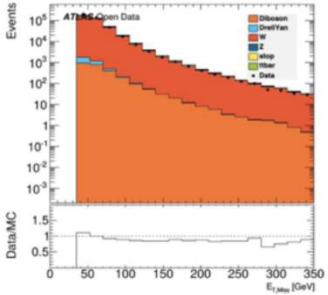




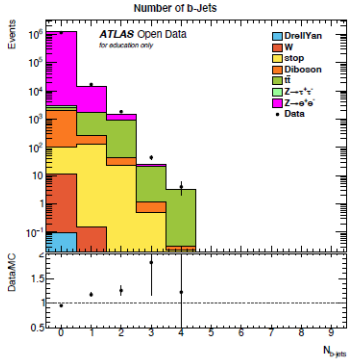
[Making ATLAS Data from CERN Accessible to the General Public (2017)]



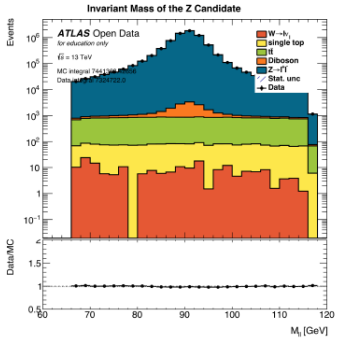
[Reconstrucción de masas invariantes de bosones del Modelo Estándar (2017)]



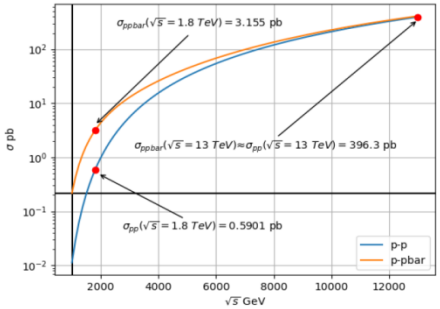
[Perspectivas y Evaluación de producción de Materia Oscura (2017)]



[Enabling Open Science with the ATLAS Open Data (2018)]



[A Contribution to ATLAS Open Data Collaboration at CERN (2019)]



[Measurements of top-quark pair production using ATLAS open data (2021)]

CMS Open Data Workshop & Hackathon 2024

Jul 29, 2024, 2:00 PM → Aug 1, 2024, 6:00 PM Europe/Zurich

IdeaSquare (CERN)

Julie Hogan (Brown University, Bethel University (US)), Kati Lassila-Perini (Helsinki Institute of Physics (FI))

Description

Since 2014, the CMS Collaboration has pioneered the release of LHC research quality data for public use by making a significant amount of these data accessible through the CERN Open Data portal. Recently in 2024, the CMS Collaboration has released a sizeable new set of 13 TeV data collected in 2016.

This workshop is the fifth in a series that started in 2020 and it aims to bridge the technical gap that usually exists between the scientific creativity of an external analyst and the nuts-and-bolts details of a full analysis with CMS open data.

All exercises will be hands-on and participants should be prepared to dive into the data right away. A set of pre-exercises are provided and required for participants so that they can make the most of the workshop.

New for this workshop are morning **hackathon segments** where users with Open Data projects in mind can take advantage of the facilitators' knowledge to jump-start their work. Tutorial sessions in the afternoons will give users at any stage a pedagogical grounding in the CMS experiment and relevant open data analysis techniques.

The tutorial segments of the workshop will be offered in hybrid mode. The hackathon working segments will be optimized for in-person participation.

Access the tutorial site here (coming in July)

Open data from CERN (each experiment having its own open data) has proven to be highly successful in outreach and education.

The tools are designed to assist students and teachers and foster educational efforts at both undergraduate and graduate levels.

I invite you to use them in your courses or mention their existence and potential use to your professors and teachers in their classes and science outreach efforts!

To start using CERN's open data, check out the tools and resources provided on the CERN Open Data portal, where access guides, analysis examples, and necessary software to process the data are offered.

opendata CERN

https://opendata.cern.ch/