

Open Data at CERN

UNIGE Open Science Course: Open Science at CERN

Pablo Saiz

22 Nov 2023

Slides taken from Tibor Šimko

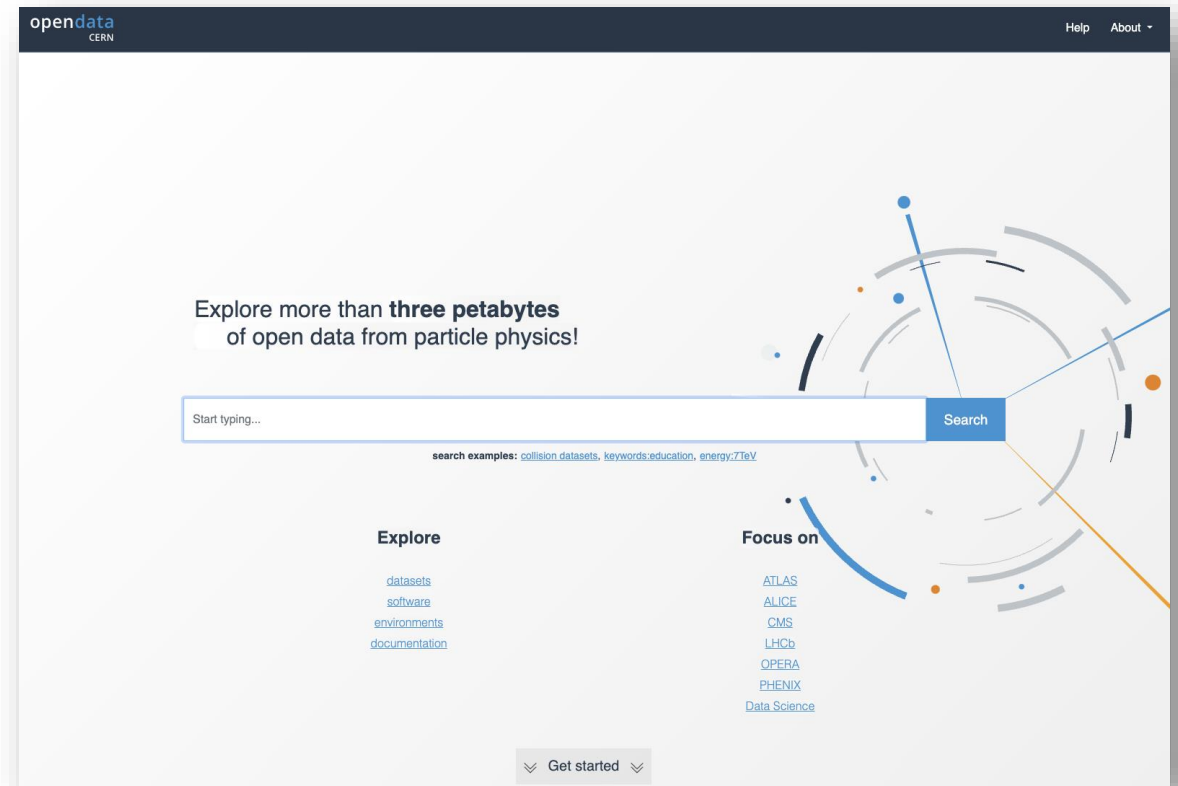
Content

- **CERN Open Data portal**
- **FAIR principle**
- **REANA**
- **Summary**

CERN Open Data (COD) Portal

- **Repository of data**
 - Launched in November 2014
- **Plenty of content**
 - Dataset
 - Collision, simulated and derived
 - Documentation
 - Glossary, tutorials, configuration, examples
 - Software
 - Frameworks, virtual machines, containers
- **Current size (Nov 2023)**
 - > 17.000 records
 - > 1.900.000 files
 - > 4.5 PB

<http://opendata.cern.ch>



Developed by CERN in collaboration with Experiments



Enables independent theoretical research

Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski (Reed Coll.), Simone Marzani (SUNY, Buffalo), Jesse Thaler (MIT, Cambridge, CTP), Aashish Tripathhee (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 60 citations

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→2 splitting function of QCD. To our knowledge, this is the first ever physics analysis based on the CMS Open Data.

Note: 7 pages, 5 figures; v2: references updated and figure formatting improved; v3: approximate version to appear in PRL

jet: energy energy: high singularity: collinear splitting function quantum chromodynamics: perturbation theory CMS parton

References (102) Figures (5)

[1] Deep inelastic e p scattering in perturbation theory
V.N. Gribov (St. Petersburg, INF), L.N. Lipatov (St. Petersburg, INF)
Sov.J.Nucl.Phys. 15 (1972) 438-450, *Yad.Fiz.* 15 (1972) 781-807

[2] Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.
Yuri L. Dokshitzer (St. Petersburg, INF)
Sov.Phys.JETP 46 (1977) 641-653, *Zh.Eksp.Theor.Fiz.* 73 (1977) 1216-1240

[76] Jet primary dataset in AOD format from RunB of 2010 (JetRun2010B-Apr2 1HeReco→1/AOD, CERN Open Data Portal DOI: CMS Collaboration Collaboration
DOI: 10.2483/OPENDATA.CMS.3577.2E99

Feedback

Over thirty papers citing CMS Open Data

Measurement of the Splitting Function in pp and Pb-Pb Collisions at $\sqrt{s_{NN}} = 5.02$ TeV

CMS Collaboration - Albert M Sirunyan (Yerevan Phys. Inst.) (Show All(2258))
Aug 30, 2017

17 pages
Published in: *Phys.Rev.Lett.* 120 (2018) 14, 142302
Published: Apr 3, 2018
e-Print: 1708.09429 [nucl-ex]
DOI: 10.1103/PhysRevLett.120.142302
Report number: CMS-HIN-16-006, CERN-EP-2017-205
Experiments: CERN-LHC-CMS
View in: HAL Science Ouverte, OSTI Information Bridge Server, CERN Document Server, ADS Abstract Service

pdf links cite claim datasets reference search 153 citations

Abstract: (APS)
Data from heavy ion collisions suggest that the evolution of a parton shower is modified by interactions with the color charges in the dense partonic medium created in these collisions, but it is not known where in the shower evolution the modifications occur. The momentum ratio of the two leading partons, resolved as two other partons and has been measured for jets with transverse momentum between 140 and 500 GeV, in pp and PbPb collisions at a center-of-mass energy of 5.02 TeV per nucleon pair. In central PbPb collisions, the splitting function indicates a more unbalanced momentum ratio, compared to peripheral PbPb and pp collisions... The measurements are compared to various predictions from event generators and analytical calculations.

Note: Replaced with the published version. Added the journal reference and the DOI. All the figures and tables can be found at <http://cms-results.web.cern.ch/cms-results/public-results/publications/HIN-16-006> (CMS Public Pages)

splitting function parton: showers parton: splitting jet: transverse momentum quark gluon: plasma charge: color particle flow anti-kt algorithm vertex: primary jet: energy Show all (21)

References (63) Figures (4)

Jet Quenching in Dense Matter

[1] Miklos Gyulassy (LBL, Berkeley), Michael Plumer (LBL, Berkeley)
Phys.Lett.B 243 (1990) 432-438 • DOI: 10.1016/0370-2693(90)91409-5

Exposing the QCD Splitting Function with CMS Open Data

[33] Andrew Larkoski (Reed Coll.), Simone Marzani (SUNY, Buffalo), Jesse Thaler (MIT, Cambridge, CTP), Aashish Tripathhee (MIT, Cambridge, CTP), Wei Xue (MIT, Cambridge, CTP)
Phys.Rev.Lett. 119 (2017) 13, 132003 • e-Print: 1704.05066 • DOI: 10.1103/PhysRevLett.119.132003

Feedback

... that the CMS collaboration cites

LHC collaboration 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 collaborators, the wider scientific community and the general public to analyze data for educational purpose and for eventual reanalysis 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 condition parameters are essential 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 constituting 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 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.

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 DPREP [1] 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. [2]. 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 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[3]. 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 RHEAT [4]. 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 [5]. 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

Approved by the CMS Collaboration Board 20th April 2014

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 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 reconstructed data and the datasets of higher abstraction generated by a analysis workflows, and finally all the ways 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. 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. Hübner, F. and Kerner, S. eds. "Data results from the ATLAS high-purity HEP survey on data preservation, re-use and open access." <http://inspirehep.net>

LHCb Data Access Policy

Reference: LHCb-PUB-2013-003

Version: 1
Issue: 1
Date: 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 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. The extent which 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 DPREP 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 not 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 unforeseen consequences arising.

page 1



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


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.

FAIR: Findable

Data and metadata should be easy to find by humans and computers

Principles:

- 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

COD Findable

Search results for 'mu primary' showing filters and results. A red circle highlights the filter sidebar.

Filters:

- Filter by type: Dataset (207), Collision (207)
- Filter by experiment: CMS (207)
- Filter by year: 2010 (26), 2011 (37), 2012 (94), 2015 (50)
- Filter by file type: aod (207), miniaod (17), raw (16), reco (36), root (276)
- Filter by collision type: pp (37)
- Filter by collision energy: 13TeV (17), 5.02TeV (33)

F4: Searchable

Dataset details for 'Mu primary dataset in AOD format from RunB of 2010'. A red circle highlights the description and system details sections.

Description
Mu primary dataset in AOD format from RunB of 2010
The list of validated runs, which must be applied to all analyses, can be found in [CMS list of validated runs CerT_136033-149442_7TeV_Apr21ReReco_Collisions10_JSON_v2.txt](#)

Dataset characteristics
32376291 events, 2979 files, 2.9 TiB in total.

System details
Recommended global tag for analysis: FT_R_42_V10A::All
Recommended release for analysis: CMSSW_4_2_8
Recommended container image for analyses is available in the following locations (see guide):

- docker.io/cmsopendata/cmssw_4_2_8-slc5_amd64_gcc434:latest
- gitlab-registry.cern.ch/cms-c-cloud/cmssw-docker-opendata/cmssw_4_2_8-slc5_amd64_gcc434:latest

F1: DOI

F2: Rich metadata

F3: Documentaion

Document titled 'CMS Guide to the CMS condition database' with sections on global tags and condition data.

What is the condition database?
This page explains the use of global tags and the condition database with the CMS Open Data.

When is the condition database needed?
Most AOD and MINIAOD physics objects in the CMS Open Data are already calibrated and ready-to-use, and no additional corrections are needed other than selection and identification criteria, which will be applied in the analysis code. Therefore, simple analyses do not need to access the condition database.

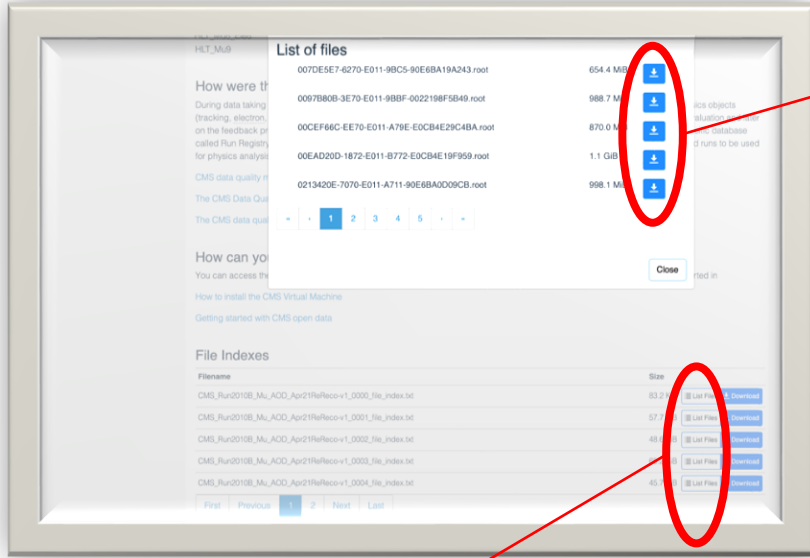
FAIR: Accessible

Once the users find the required data, they need to know how it can be access

Principles:

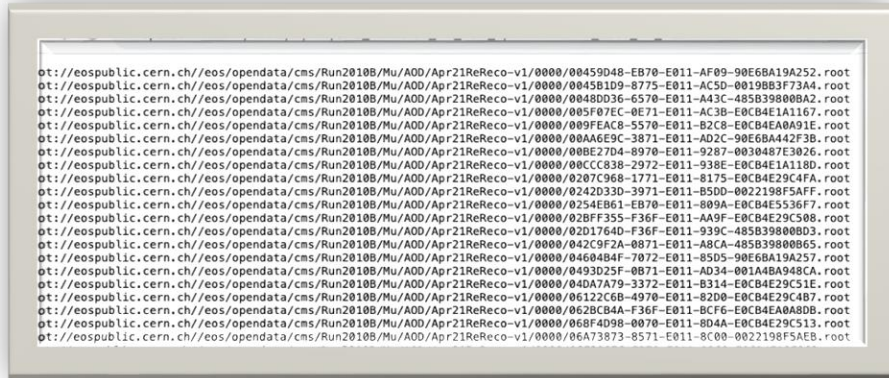
- A1: (Meta)data are retrievable by their identifier using a standardised communications protocol
 - A1.1 The protocol is open, free, and universally implementable
 - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

COD: Accessible



A1: Retriavable

Big datasets



<https://github.com/cernopendata/cernopendata-client>

```
[bash-5.1$ cernopendata-client --help
Usage: cernopendata-client [OPTIONS] COMMAND [ARGS]...

Command-line client for interacting with CERN Open Data portal.

Options:
  --help Show this message and exit.

Commands:
  download-files      Download data files belonging to a record.
  get-file-locations  Get a list of data file locations of a record.
  get-metadata        Get metadata content of a record.
  list-directory      List contents of a EOSPUBLIC Open Data directory.
  verify-files        Verify downloaded data file integrity.
  version             Return cernopendata-client version.
bash-5.1$
```

```
[bash-5.1$ cernopendata-client get-file-locations --recid 14
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00459D48-EB70-E011-AF09-90E6BA19A252.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0045B1D9-8775-E011-AC5D-0019BB3F73A4.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0048DD36-6570-E011-A43C-485B39800BA2.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/005F07EC-0E71-E011-AC3B-E0CB4E1A1167.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/009FEACB-5570-E011-B2CB-E0CB4E0A091E.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00AE69C-3871-E011-AD2C-90E6BA442F3B.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00BE27D4-8970-E011-9287-0030487E3026.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0242D33D-3971-E011-B5DD-0022198F5AFF.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0207C968-1771-E011-8175-E0CB4E29C4FA.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0254EB61-EB70-E011-809A-E0CB4E5536F7.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/02BF355-F36F-E011-AA9F-E0CB4E29C508.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/02D1764D-F36F-E011-939C-485B39800BD3.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/042C9F2A-0871-E011-A8CA-485B39800B65.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/04604B4F-7072-E011-85D5-90E6BA19A257.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/04604B4F-7072-E011-85D5-90E6BA19A257.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/047A79-3372-E011-B314-E0CB4E29C51E.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0493D25F-0B71-E011-AD34-001A4BA948CA.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/06122C6B-4970-E011-82D0-E0CB4E29C487.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/0628CB4A-F36F-E011-BCF6-E0CB4E0A08DB.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/068F4D98-0070-E011-8D4A-E0CB4E29C513.root
http://opendata.cern.ch/eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/06A73873-8571-E011-8C00-0022198F5AEB.root
```

FAIR: Interoperable

Data usually need to be integrated with other data, and with applications/workflows for analysis, storage and processing.

- Principles:
 - 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

COD: Interoperable

Filter by file type

- C 4
- DST 15
- MDST 12
- aod 207
- aodsim 850
- cc 12
- csv 939
- db 18
- docx 1
- fevtdebughit 1
- gen-sim 5
- gen-sim-digi-raw 1
- gen-sim-reco 225
- gz 18
- h5 4
- html 7
- lg 95
- ipynb 2
- jpg 1
- json 14
- m4v 1
- miniaod 17
- miniaodsim 7131
- nanoaod 18
- ova 2
- pdf 19
- png 3
- py 1328
- raw 16
- reco 36
- root 8593
- tar 2
- tar.gz 1
- tgz 2
- txt 41
- xls 1
- xml 7
- zip 38

Filter by category

- Heavy-Ion Physics 219
- B physics and Quarkonia 72
- Beyond 2 Generations 239
- Exotica 4469
 - Dark Matter 600
 - Excited Fermions 207
 - Extra Dimensions 585
 - Gravitons 880
 - Heavy Fermions, Heavy Righ-Han... 493
 - Heavy Gauge Bosons 1017
 - Leptoquarks 443
 - Miscellaneous 244
- Heavy-Ion Physics 1
- Higgs Physics 1618
 - Beyond Standard Model 553
 - Standard Model 1065
- Physics Modelling 53
- Standard Model Physics 1044
 - Drell-Yan 112
 - ElectroWeak 377
 - Forward and Small-x QCD Physics 25
 - Minimum Bias 32
 - QCD 251
 - Top physics 247
- Supersymmetry 491

I1: Multiple data formats and common classifications

Dataset semantics table:

Variable	Type	Description
hit_global_x	std::vector<float>	global x position of the RecHit
hit_global_y	std::vector<float>	global y position of the RecHit
hit_global_z	std::vector<float>	global z position of the RecHit
hit_local_x	std::vector<float>	x pos. of the hit in the local sensor coordinate
hit_local_y	std::vector<float>	y pos. of the hit in the local sensor coordinate
hit_local_x_error	std::vector<float>	x error in the local sensor coordinate
hit_local_y_error	std::vector<float>	y error in the local sensor coordinate
hit_sub_det	std::vector<unsigned int>	subdetector generating the hit [1 PixelBarrel, 2 PixelEndcap, 3 TIB, 4 TID, 5 TOB, 6 TEC]
hit_layer	std::vector<unsigned int>	layer/disk of the subdetector generating the hit
hit_type	std::vector<unsigned int>	type of strip hit [0 Pixel hit, 1 rphiRecHit, 2 stereoRecHit, 3 rphiRecHitUnmatched, 4 stereoRecHitUnmatched]
hit_simtrack_id	std::vector<int>	ID number of the sim track matched to the hit
hit_simtrack_index	std::vector<unsigned int>	index of the sim track matched to the hit
hit_simtrack_match	std::vector<bool>	is the hit matched to a sim track?

I2: Semantic descriptions

Analysis of the di-muon spectrum using data from the CMS detector taken in 2012

Description

This analysis takes data from the CMS experiment recorded in 2012 during Run B and C and extracts the di-muon spectrum. The di-muon spectrum is computed from the data by calculating the invariant mass of muon pairs with opposite charge. In the resulting plot, you are able to rediscover particle resonances in a wide energy range from the η meson at about 548 MeV up to the Z boson at about 91 GeV.

The analysis code opens an interactive plot, which allows to zoom and navigate in the spectrum. Note that the bump at 30 GeV is not a resonance but an effect of the data taking due to the used trigger. The technical description of the dataset can be found in the respective record linked below.

The result of this analysis can be compared with an official result of the CMS collaboration using data taken in 2010, see the plots below:

Use with

[DoubleMuParked dataset from 2012 in NanoAOD format reduced on muons](#)

Related items

[DoubleMuParked dataset from 2012 in NanoAOD format reduced on muons](#)

[/DoubleMuParked/Run2012B-22Jan2013-v1/AOD](#)

[/DoubleMuParked/Run2012C-22Jan2013-v1/AOD](#)

I3: Fully qualified references

FAIR: Reusable

Metadata and data should be well-described so that they can be replicated and/or combined in different settings

- Principles:
 - 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

COD: Reusable

Simulated dataset BulkGravTohhTohhbbbb_narrow_M-4500_13TeV-madgraph in MINIAODSIM format for 2016 collision data

How were these data generated?

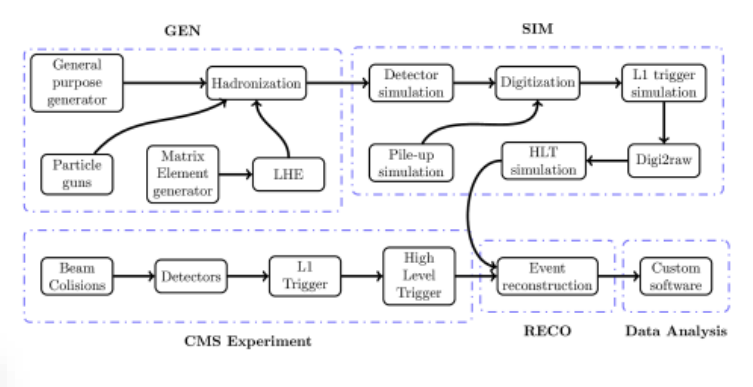
These data were generated in several steps (see also CMS Monte Carlo production overview):

Step LHE
Release: CMSSW_7_1_16
Output dataset: /BulkGravTohhTohhbbbb_narrow_M-4500_13TeV-madgraph/RunI/Summer16MiniAODv2-PU/Moriond17_80X_mcRun2_asymptotic_2016_TracheIV_v6-v1/MINIAODSIM, CMS Collaboration
Note: To get the exact generator parameters, please see Finding the generator parameters.

Step SIM
Release: CMSSW_7_1_20
Configuration file for SIM (link)

Disclaimer
The open data are released under the Creative Commons CC0 waiver. Neither CMS nor CERN endorse any works, scientific or otherwise, produced with these data. All releases will have a unique DOI that you are requested to cite in any applications or publications.

R1.1: Data usage license



R1.2: Provenance

Reuse/reproduce: Can we reproduce analysis? sample?

SingleElectron primary dataset sample in RAW format from RunA of 2011 (from /SingleElectron/Run2011A-v1/RAW)
Cite as: CMS collaboration (2019). SingleElectron primary dataset sample in RAW format from RunA of 2011 (from /SingleElectron/Run2011A-v1/RAW). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.P87Z.TXTV

Dataset characteristics
2064298 events, 395.2 GiB in total.

System details
Recommended global tag for reconstruction: FT_53_LV5_A
Recommended release for reconstruction: CMSSW_5_3_32
Recommended container image for analyses is available in:
• docker.io/cmsopendata/cmsw_5_3_32-slc6_amd64
• gitlab-registry.cern.ch/cms-cloud/cmsw-docker

How were these data selected?
Events stored in this primary dataset were selected because of the presence of at least one high-energy electron in the event.

Data taking / HLT
The collision data were assigned to different RAW datasets using the following HLT configuration.

SingleElectron primary dataset in AOD format from RunA of 2011 (/SingleElectron/Run2011A-12Oct2013-v1/AOD)
Cite as: CMS collaboration (2016). SingleElectron primary dataset in AOD format from RunA of 2011 (/SingleElectron/Run2011A-12Oct2013-v1/AOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.P87Z.TXTV

Dataset characteristics
41709195 events, 1542 files, 5.2 TiB in total.

System details
Recommended global tag for analysis: FT_53_LV5_AN1
Recommended release for analysis: CMSSW_5_3_32
Recommended container image for analyses is available in the following locations (see guide):
• docker.io/cmsopendata/cmsw_5_3_32-slc6_amd64_gcc472:latest
• gitlab-registry.cern.ch/cms-cloud/cmsw-docker-opendata/cmsw_5_3_32-slc6_amd64_gcc472:latest

How were these data selected?
Events stored in this primary dataset were selected because of the presence of at least one high-energy electron in the event.

Data taking / HLT
The collision data were assigned to different RAW datasets using the following HLT configuration.

Search Docker Hub [3K+K] Explore Pricing Sign In Sign up

CMS Collaboration
Sponsored OSS
Community Organization CMS Collaboration Joined July 10, 2019

Repositories
Displaying 1 to 14 repositories

- cmsopendata/cmsw_5_3_20-slc6_amd64_gcc472** Sponsored OSS · 19 · 0
By CMS Collaboration · Updated 3 months ago
- cmsopendata/cmsw_10_6_30-slc7_amd64_gcc700** Sponsored OSS · 27 · 0
By CMS Collaboration · Updated 4 months ago
- cmsopendata/cmsw_7_6_7-slc6_amd64_gcc493** Sponsored OSS · 1.3K · 1
By CMS Collaboration · Updated 8 months ago

Reuse/reproduce: Can we reproduce analysis?

<https://pubmed.ncbi.nlm.nih.gov/22675527/>

> PLoS One. 2012;7(6):e38234. doi: 10.1371/journal.pone.0038234. Epub 2012 Jun 1.

The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements

Ed H B M Gronenschild¹, Petra Habets, Heidi I L Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

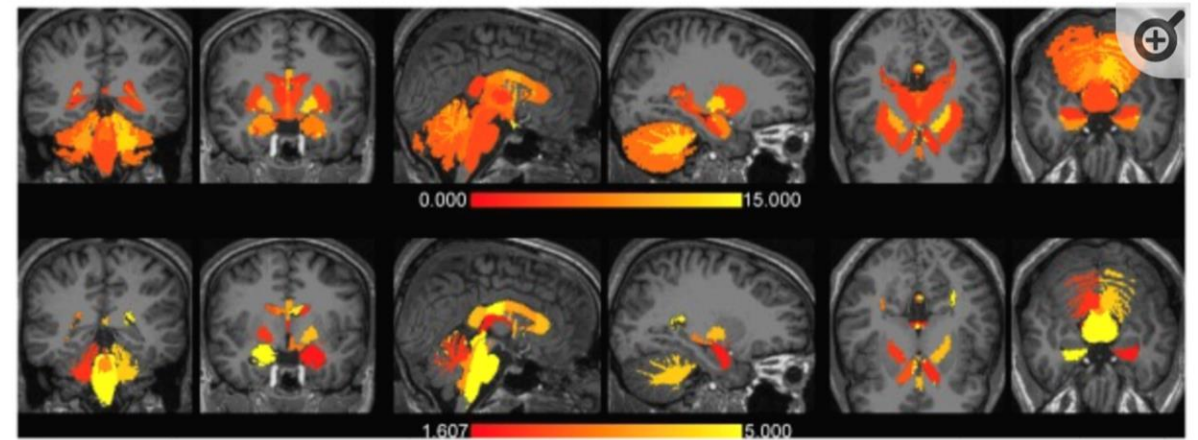
Affiliations + expand

PMID: 22675527 PMCID: PMC3365894 DOI: 10.1371/journal.pone.0038234

[Free PMC article](#)

Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average 8.8 ± 6.6% (volume) and 2.8 ± 1.3% (thickness).



Software changes (Freesurfer 4.3.1, 4.5.0, 5.0.0):
8.8±6.6% (volume) and 2.8±1.3% (thickness)

Operating system changes (macOS 10.5, 10.6):
about factor two smaller

Four pillars of reusable computational research

I. Input Data:

- I. Input Files and parameters

II. Analysis code:

- I. User code
- II. Software frameworks

III. Computing Environment:

- I. Operating system
- II. Databases

IV. Computational recipes:

- I. Extra shell commands steps
- II. Notebooks and workflows

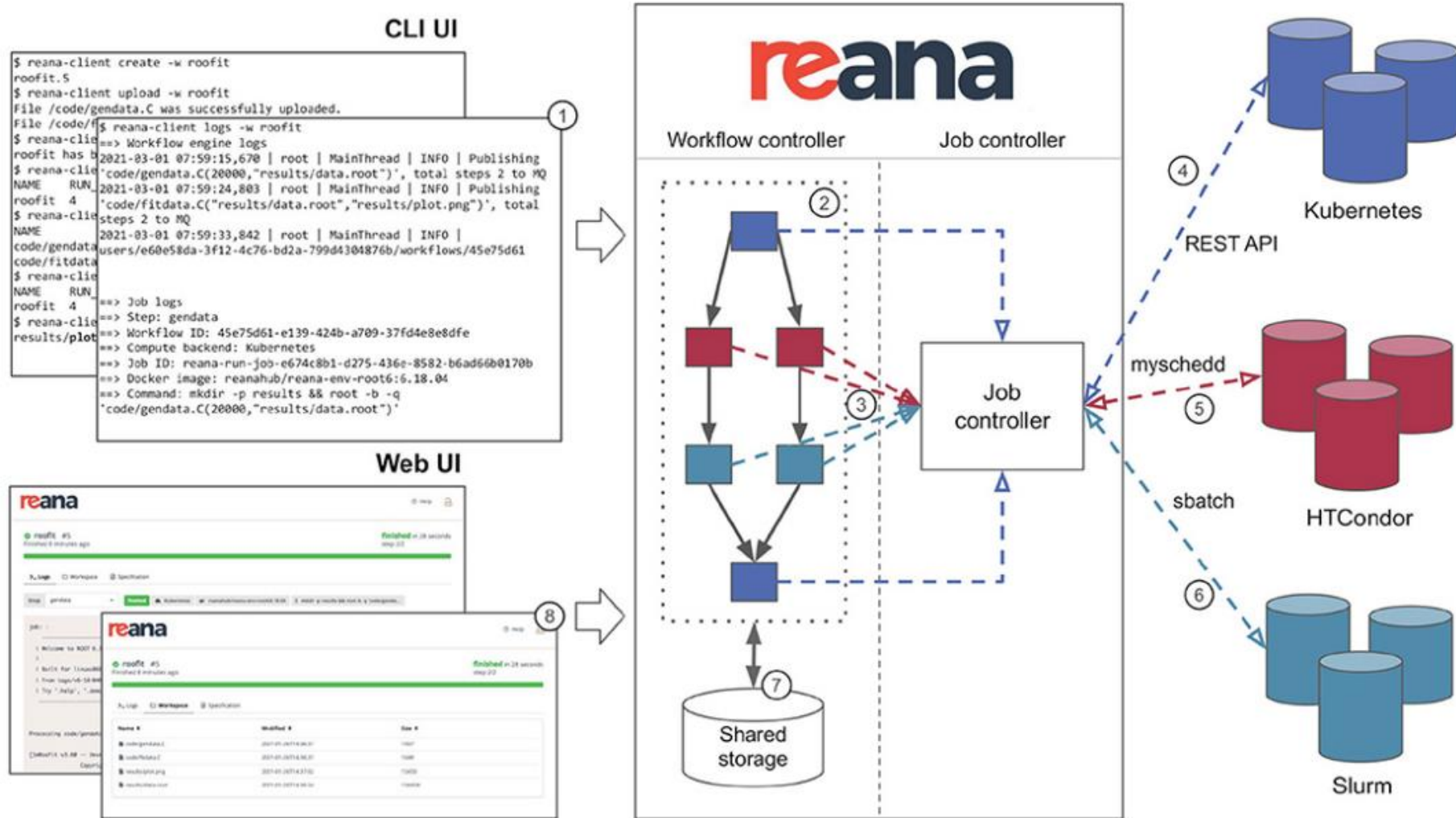
REANA: Reusable analysis

<http://reana.io>

Deploy and run
containerised
workflows on
compute clouds

The screenshot shows the REANA website homepage. At the top left is the 'reana' logo. To the right is a navigation menu with links: Home, Examples, Get Started, Documentation, News, Roadmap, Contact, and Blog. The main heading is 'reana' in a large, bold font, with 're' in red and 'ana' in dark blue. Below this is the tagline 'Reproducible research data analysis platform'. The page is divided into four columns, each representing a key feature: Flexible, Scalable, Reusable, and Free. Each column contains a brief description and logos of supported technologies. The 'Flexible' column mentions 'Run many computational workflow engines' and shows logos for 'COMMON WORKFLOW LANGUAGE', 'S', and a blue triangle. The 'Scalable' column mentions 'Support for remote compute clouds' and shows logos for 'kubernetes', 'HTCondor High Throughput Computing', and 'slurm workload manager'. The 'Reusable' column mentions 'Containerise once, reuse elsewhere. Cloud-native.' and shows logos for Docker and a circular 'S' logo. The 'Free' column mentions 'Free Software. MIT licence. Made with ❤️ at CERN.' and shows the CERN logo.

REANA in a nutshell



SUMMARY: Open is not enough

Findable
 Accessible
 Interoperable
 Reusable

Data
 +
 Code
 +
 Environment
 +
 Workflow
 =====
 Reusability

nature physics PERSPECTIVE
 https://doi.org/10.1038/s41567-018-0342-2 OPEN

Corrected: Publisher Correction

Open is not enough

Xiaoli Chen^{1,2}, Sünje Dallmeier-Tiessen^{1*}, Robin Dasler^{1,3}, Sebastian Feger^{1,3}, Pamfilos Fokianos⁴, Jose Benito Gonzalez⁵, Harri Hirvonsalo^{4,6,7}, Dinos Kousidis⁸, Artemis Lavasa⁹, Salvatore Mele¹, Diego Rodriguez Rodriguez², Tibor Šimko^{1*}, Tim Smith¹, Ana Trisovic^{1,5*}, Anna Trzcinska¹, Ioannis Tsanaktsidis¹, Markus Zimmermann¹, Kyle Cramer⁶, Lukas Heinrich⁶, Gordon Watts⁶, Michael Hildreth⁶, Lara Lloret Iglesias⁶, Kati Lassila-Perini⁶ and Sebastian Neubert¹⁰

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 work-

flows 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 Goble⁴ and Lorena A. Barba⁵ 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 analysis themselves are intrinsically complex due to the large data volume and algorithms involved⁶. In addition, the analysis 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.

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

NATURE PHYSICS | VOL 15 | FEBRUARY 2019 | 113–119 | www.nature.com/naturephysics 113

PERSPECTIVE NATURE PHYSICS

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 analysis. 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), Dave 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 lab
Reuse	Transfer	Different experiment

potentially leading to a loss of knowledge concerning how the results were obtained⁷. 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 processes⁸. Because these reviews often involve the input of the entire collaboration, where the level of crosschecking is extensive, the measurements are considered trustworthy.

NATURE PHYSICS | VOL 15 | FEBRUARY 2019 | 113–119 | www.nature.com/naturephysics 114

<https://www.nature.com/articles/s41567-018-0342-2.pdf>





home.cern