

CERN Open Data portal: preserve to reuse

Tibor Šimko

CERN

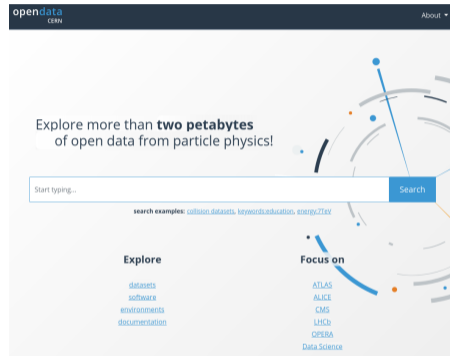
*Unveiling hidden Physics Beyond the Standard Model at the LHC
1–3 March 2021*

<https://indico.tlabs.ac.za/event/100>

CERN Open Data portal

CERN Open Data portal

- ▶ launched in November 2014
- ▶ rich content
 - ▶ collision and simulated datasets for research
 - ▶ derived datasets for education
 - ▶ configuration files and documentation
 - ▶ virtual machines and container images
 - ▶ software tools and analysis examples
- ▶ total size in March 2021
 - ▶ over 7'600 bibliographic records
 - ▶ over 900'000 files
 - ▶ over 2.4 petabytes

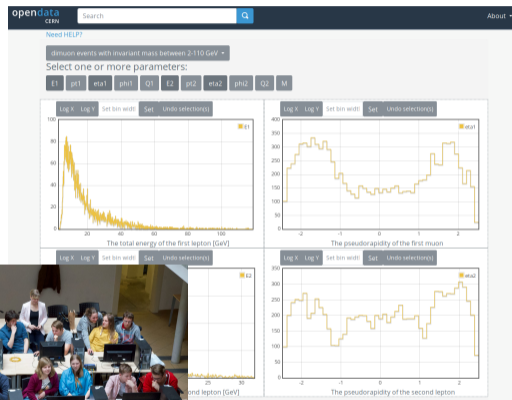
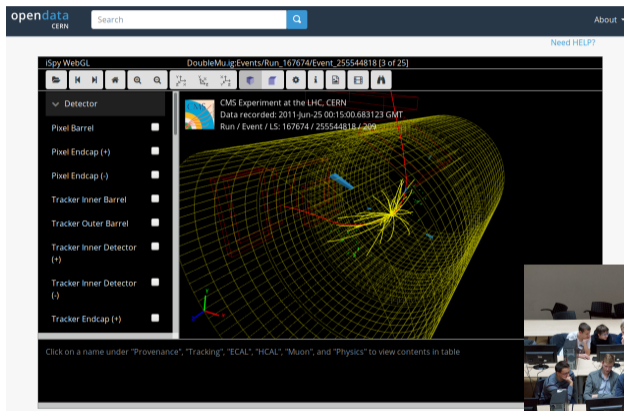


<https://opendata.cern.ch>

Developed by CERN in close collaboration with Experiments

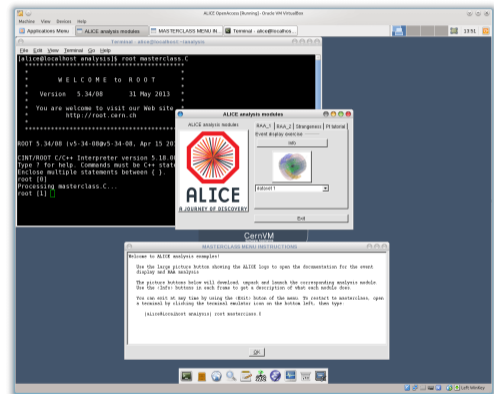


Education-oriented use cases



Interactive event display and histogramming for derived datasets

Research-oriented use cases



Run CernVM Virtual Machines

open data
CERN

About

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

Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Alzuddin

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Alzuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS_JK88.RR4Z

Software
Analysis
CMS
Accelerator CERN/LHC

Description

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

The published reference plot which is being approximated in this example is https://inspirehep.net/record/1124338/files/H41_mass_3.png. Other Higgs final states (e.g. Higgs to two photons), which were also part of the same CMS paper and strongly contributed to the Higgs boson discovery, are not covered by this example.

The example consists of different levels of complexity. The highest level of this example addresses users who feel they have at least some minimal understanding of the content of this paper and of the meaning of this reference plot, which can be reached via (separate) educational exerc with the linux op

Use with

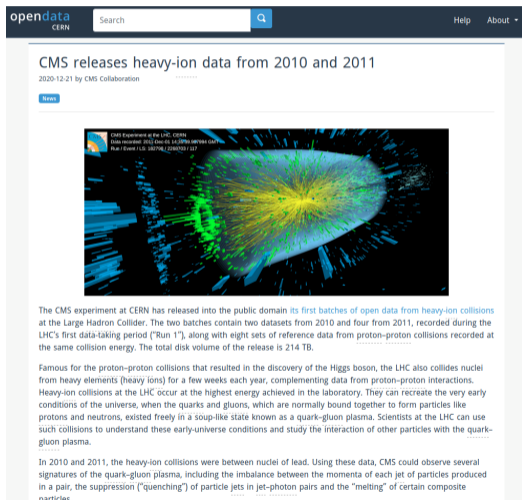
The example uses publication due to but not identical in many later CM

`/DoubleElectron/`
`/DoubleMu/Run2`

re original again close to, ly as they are,

Run simplified research-level analysis examples

Latest research-level content news



The screenshot shows the OpenData CERN website. At the top, there is a search bar and navigation links for 'Help' and 'About'. The main content area features a news article with the title 'CMS releases heavy-ion data from 2010 and 2011' and a sub-headline '2020-12-21 by CMS Collaboration'. Below the text is a large, colorful visualization of a heavy-ion collision event, showing a dense, multi-colored (yellow, green, blue) particle shower. The article text describes the release of the first batches of open data from heavy-ion collisions at the LHC, mentioning the datasets from 2010 and 2011, and the 'Run 1' data-taking period. It also explains the scientific goals of heavy-ion collisions, such as recreating early universe conditions and studying quark-gluon plasma.

- ▶ **August 2020** CMS released fifth batch of new open data. All proton-proton collision data recorded in 2010–11 and half of 2012 are available throughout the portal.

<https://opendata.cern.ch/docs/cms-completes-2010-2011-pp-data>

- ▶ **October 2020** CMS run a CMS Open Data Workshop for Theorists at the LPC.

<https://indico.cern.ch/event/882586/>

- ▶ **December 2020** CMS released first 2010–11 heavy-ion data samples and reference proton-proton datasets (214 TB).

<https://opendata.cern.ch/docs/cms-releases-heavy-ion-data>

Enables independent theoretical research



Welcome to **INSPIRE**, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEP-NAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

reference:10.7483/OPENDATA.CMS Brief format Search Save Search Advanced Search
[View "Phys.Rev.Lett." > more](#) Search the new INSPIRE

Sort by: Display results:
latest first desc times cited 25 results single list

No exact match found for 10.7483/OPENDATA.CMS, using 10 7483 OPENDATA CMS instead...

HEP 35 records found 1 - 25 ▶ jump to record: 1 Search took 0.12 seconds.

1. Exposing the QCD Splitting Function with CMS Open Data

⁽²⁰⁾ Andrew Larkoski (Reed Coll.), Simone Marzani (SUNY, Buffalo), Jesse Thaler, Aashish Tripathhee, Wei Xue (MIT, Cambridge, CTP), Apr 17, 2017, 7 pp.

Published in *Phys.Rev.Lett.* **119** (2017) no.13, 132003

MIT-CTP-4891

DOI: [10.1103/PhysRevLett.119.132003](https://doi.org/10.1103/PhysRevLett.119.132003)

e-PRINT: [arXiv:1704.05066](https://arxiv.org/abs/1704.05066) [hep-ph] | PDF

References | BibTeX | LaTeX (US) | LaTeX (EU) | HarvMatic | EndNote
ADS Abstract Service

Detailed record - Cited by 39 records

2. Fast and Accurate Simulation of Particle Detectors Using Generative Adversarial Networks

⁽³³⁾ Pasquale Musella (ETH, Zurich (main)), Francesco Pandolfi (INFN, Rome), May 2, 2018, 8 pp.

Published in *Comput.Softw.Big Sci.* **2** (2018) no.1, 8

DOI: [10.1007/s41781-018-0012-y](https://doi.org/10.1007/s41781-018-0012-y)

e-PRINT: [arXiv:1805.09850](https://arxiv.org/abs/1805.09850) [hep-ex] | PDF

References | BibTeX | LaTeX (US) | LaTeX (EU) | HarvMatic | EndNote
ADS Abstract Service

Detailed record - Cited by 33 records

3. Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2

⁽¹⁸⁾ LHC Reinterpretation Forum Collaboration (Waleed Abdallah (Harish-Chandra Res. Inst. & Carro U) et al.), Mar 19, 2020, 58 pp.

Published in *SciPost Phys.* **9** (2020) no.2, 022

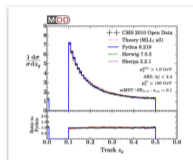
CERN-LPCC-2020-001, FERMILAB-FN-1098-CMS-T, Imperial/HEP/2020/RF/01

DOI: [10.21468/SciPostPhys.9.2.022](https://doi.org/10.21468/SciPostPhys.9.2.022)

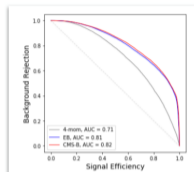
e-PRINT: [arXiv:2003.07868](https://arxiv.org/abs/2003.07868) [hep-ph] | PDF

References | BibTeX | LaTeX (US) | LaTeX (EU) | HarvMatic | EndNote
CERN Document Server: ADS Abstract Service: OSTI.gov Server: Link to Fermilab Library Server (fulltext available): Link to Fulltext from Publisher: Link to Fulltext

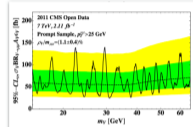
Detailed record - Cited by 19 records



arXiv:1704.05066



arXiv:1807.11916



arXiv:1902.04222

Searches, QCD jet studies, Machine Learning...

A measurement of the z_j 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](https://doi.org/10.1103/PhysRevLett.119.132003), [arXiv:1704.05066](https://arxiv.org/abs/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](https://doi.org/10.1103/PhysRevD.96.074003), [arXiv:1704.05842](https://arxiv.org/abs/1704.05842).

... that the CMS collaboration cites!

Over thirty papers citing CMS open data

arXiv:1708.09429v2

A FAIRy tale

FAIR guiding principles for scientific data management

scientific **data**

[Explore Content](#) ▾ [Journal Information](#) ▾ [Publish With Us](#) ▾

[nature](#) > [scientific data](#) > [comment](#) > [article](#)


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

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] [Barend Mons](#) 

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

180k [Accesses](#) | **2306** [Citations](#) | **1797** [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.

- ▶ Findable
- ▶ Accessible
- ▶ Interoperable
- ▶ Reusable

FAIR principles: F is for Findable

The first step in (re)using data is to find them.

Metadata and data should be easy to find for both humans and computers.

Machine-readable metadata are essential for automatic discovery of datasets and services.

- ▶ 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 they describe
- ▶ F4. (Meta)data are registered or indexed in a searchable resource

Dataset information: persistent identifiers & machine readability

The screenshot shows the OpenData CERN website interface. At the top, there is a search bar and an 'About' link. The main content area displays the title 'Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD)' and the source 'CMS collaboration'. Below this, there are tabs for 'Dataset', 'Collaboration', 'CMS', 'Collision energy /TeV', 'Accelerator/CERN/LHC', and 'Parent Dataset: /Mu/Run2010B-v1/RAW'. The 'Description' section states 'Mu primary dataset in AOD format from RunB of 2010'. The 'Notes' section explains that the dataset contains all runs from 2010 RunB and provides a link to a CMS list of validated runs. The 'Related Datasets' section lists '/Mu/Run2010B-v1/RAW'. The 'Characteristics' section shows 'Dataset: 32376291 events 2979 files 3.2 TB in total'. The 'System Details' section includes a global tag 'FT_R_42_V10A:All' and a recommended release for analysis 'CMS5W_A_2_1_patch'.

```
{
  "created": "2020-12-21T10:16:53.741747+00:00",
  "id": 14,
  "metadata": {
    "schema": "http://opendata.cern.ch/schema/records/record-v1.0.0.json",
    "abstract": {
      "description": "p>p>Mu primary dataset in AOD format from RunB of 2010/p> <p>This dataset contains all runs from 2010 RunB. The list of validated runs, which must be applied to all analyses, can be found <a href='\"#\">here</a>."
    }
  },
  "links": {
    "record": "1000"
  }
},
{
  "accelerator": "CERN-LHC",
  "collaboration": {
    "name": "CMS collaboration",
    "recid": "459"
  },
  "collections": [
    "CMS-Primary-Datasets"
  ],
  "collision_information": {
    "energy": "13TeV",
    "type": "pp"
  },
  "control_number": "14",
  "date_created": {
    "2010"
  },
  "date_published": "2014",
  "date_reprocessed": "2011",
  "distribution": {
    "formats": [
      "root",
      "aod"
    ]
  },
  "number_events": 32376291,
  "number_files": 2979,
  "size": 326262517610
},
{
  "doi": "10.7483/OPENDATA.CMS.B8MR.C4A2",
  "experiment": "CMS",
  "files": [],
  "index_files": [
    {
      "checksum": "adler32:85e137b0",
      "filename": "CMS_Run2010B_Mu_AOD_Apr21ReReco-v1_0000_file_index.json",
      "size": 180230,
      "url_http": "http://opendata.cern.ch/record/14/files/CMS_Run2010B_Mu_AOD_Apr21ReReco-v1_0000_file_index.json",
      "url_root": "root://eospublic.cern.ch/eosopendata/cms/Run2010B/MuAOD/Apr21ReReco-v1/1file-Indexes/CMS_Run2010B_Mu_AOD_Apr21ReReco-v1_0000_file_index.json"
    },
    {
      "checksum": "adler32:1869e093",
      "filename": "CMS_Run2010B_Mu_AOD_Apr21ReReco-v1_0000_file_index.txt",
      "size": 85184,
      "url_http": "http://opendata.cern.ch/record/14/files/CMS_Run2010B_Mu_AOD_Apr21ReReco-v1_0000_file_index.txt"
    }
  ]
}
```

<https://opendata.cern.ch/record/14> \equiv <http://doi.org/10.7483/OPENDATA.CMS.B8MR.C4A2>

Each dataset is identified by a “record ID” and optionally minted with a DOI

Dataset information: rich context description & machine readability

How were these data selected?

There are four categories of triggers in the Mu dataset (with significant overlaps):

-70% inclusive single muon triggers with varying trigger pt threshold 3,5,7,9,11,13,15,17,19,21 GeV plus a few with loosened quality cuts.

-20% isolated single muon triggers with varying trigger pt threshold 9,11,13,15,17 GeV.

-10% inclusive dimuon triggers with varying trigger pt threshold 3,5 GeV plus one Z->mumu trigger with loosened quality cuts.

-20% combinations of muon triggers with various pt thresholds 3,5,7,8,9,11 GeV with some EM/e/gamma or hadronic/jet energy deposit with thresholds 6-100 GeV.

How were these data validated?

During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and physics objects (tracking, electron, muon, photon, jet and MET) show the expected performance.

Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:

[CMS data quality monitoring: Systems and experiences](#)

[The CMS Data Quality Monitoring software experience and future improvements](#)

[The CMS data quality monitoring software: experience and future prospects](#)

How can you use these data?

You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in:

[How to install the CMS Virtual Machine](#)

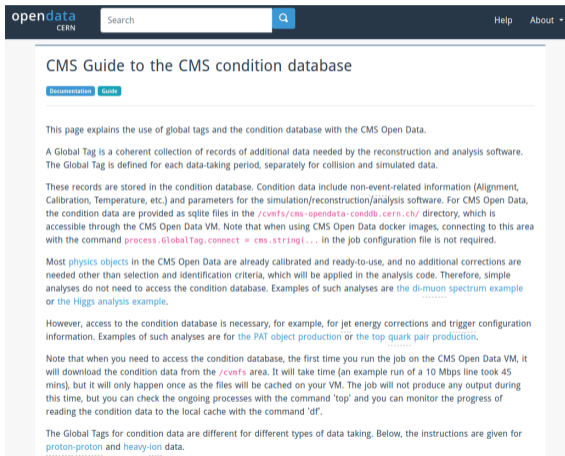
[Getting started with CMS open data](#)

```
  }
  }
  "run_period": {
    "run2010B"
  }
  "system_details": {
    "global_tag": "FT_6_42_V10A:ATL",
    "release": "CMSDK_4_2_8"
  },
  "title": "Mu/Run2010B-Apr11ReReco-v1/AOD",
  "title_additional": "Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr11ReReco-v1/AOD)",
  "type": {
    "primary": "Dataset",
    "secondary": [
      "Collision"
    ]
  },
  "usage": {
    "description": "You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in",
    "links": [
      {
        "description": "How to install the CMS Virtual Machine",
        "url": "/docs/cms-virtual-machine-2010"
      },
      {
        "description": "Getting started with CMS open data",
        "url": "/docs/cms-getting-started-2010"
      }
    ]
  },
  "validation": {
    "description": "During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and physics objects (tracking, electron, muon, photon, jet and MET) show the expected performance. Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:"
  },
  "links": [
    {
      "description": "CMS data quality monitoring: Systems and experiences",
      "url": "http://topscience.fnpp.org/1742-6596/219/7/072020/pdf/1742-6596_219_7_072020.pdf"
    },
    {
      "description": "The CMS Data Quality Monitoring software experience and future improvements",
      "url": "http://cds.cern.ch/record/1638999/files/CR2013_410.pdf"
    },
    {
      "description": "The CMS data quality monitoring software: experience and future prospects",
      "url": "http://topscience.fnpp.org/1742-6596/513/3/032024/pdf/1742-6596_513_3_032024.pdf"
    }
  ]
},
"updated": "2020-12-21T10:56:54.170861+00:00"
}
```

<https://opendata.cern.ch/record/14> \equiv <http://doi.org/10.7483/OPENDATA.CMS.B8MR.C4A2>

Rich curated context information on data selection, validation, and use

Associated documentation and guides



The screenshot shows the top navigation bar of the OpenData CERN website with a search bar and 'Help' and 'About' links. The main heading is 'CMS Guide to the CMS condition database'. Below the heading are two tabs: 'Documentation' and 'Guide'. The text explains the use of global tags and the condition database with the CMS Open Data. It defines a Global Tag as a coherent collection of records of additional data needed by the reconstruction and analysis software. It notes that records are stored in the condition database and include non-event-related information. It provides instructions on how to access the condition database through the CMS Open Data VM and how to use the `process.GlobalTag.connect` command in the job configuration file. It also mentions that physics objects are already calibrated and ready-to-use, and provides examples of analyses. Finally, it notes that access to the condition database is necessary for jet energy corrections and trigger configuration information, and provides instructions on how to download the condition data from the `/cvnfs` area.

Proton-proton data

For 2010 collision data, the global tag available in the `/cvnfs` area is `FT_R_42_V10A`. When using the "CMS-OpenData-1.1.2" VM or a higher version, it is recommended reading the condition data from there. First, set the symbolic links:

```
ln -sf /cvnfs/cms-opendata-conddb.cern.ch/FT_R_42_V10A FT_R_42_V18A
ln -sf /cvnfs/cms-opendata-conddb.cern.ch/FT_R_42_V10A.db FT_R_42_V18A.db
```

Then, define the correct set of condition data by mentioning the Global Tag in the configuration file of the job.

```
#globaltag
process.GlobalTag.connect = cms.string('sqlite_file:/cvnfs/cms-opendata-conddb.cern.ch/FT_R_42_V18A.db')
process.GlobalTag.globaltag = 'FT_R_42_V18A::All'
```

Note that **this only works in the "CMS-OpenData-1.1.2" or a higher version** of the 2010 CMS Open Data VM.

For 2010 Montecarlo data, the global tag is `START42_V17B`. To access the condition database, first, set the symbolic links:

```
ln -sf /cvnfs/cms-opendata-conddb.cern.ch/START42_V17B START42_V17B
ln -sf /cvnfs/cms-opendata-conddb.cern.ch/START42_V17B.db START42_V17B.db
```

Then, define the correct set of condition data by mentioning the Global Tag in the configuration file of the job.

```
#globaltag for 2010 MC
process.GlobalTag.connect = cms.string('sqlite_file:/cvnfs/cms-opendata-conddb.cern.ch/START42_V17B.db')
process.GlobalTag.globaltag = 'START42_V17B::All'
```

Note that **this only works in the "CMS-OpenData-1.1.2" or a higher version** of the 2010 CMS Open Data VM.

For 2011 collision data, the global tag is `FT_53_LV5_AN1`. To access the condition database, first, set the symbolic links:

```
ln -sf /cvnfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA FT_53_LV5_AN1
ln -sf /cvnfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA.db FT_53_LV5_AN1_RUNA.db
```

Make sure the `cms-opendata-conddb.cern.ch` directory has actually expanded in your VM. One way of doing this is executing:

```
ls -l
ls -l /cvnfs/
```

A detailed guide to CMS global tags and condition database

Information discovery

The screenshot shows the OpenData CERN search interface. At the top, there is a search bar and navigation links for 'Help' and 'About'. Below the search bar, there are filters for 'Include on-demand datasets' and 'Filter by type'. The 'Filter by type' section is expanded to show 'Dataset' with 2199 results. Under 'Dataset', there are sub-filters for 'Collision' (163), 'Derived' (1111), and 'Simulated' (625). The 'Documentation' section is also expanded, showing various document types like 'About', 'Activities', 'Authors', 'Guide', 'Help', 'Policy', 'Report', 'Environment', 'Condition', 'VM', 'Validation', 'Glossary', 'News', 'Software', and 'Supplementaries'. The main search results area shows 5109 results, sorted by 'Most recent' in 'asc.' order, displaying 'detailed' information for '20 results'. Three results are visible, each with a title, a link to the dataset, a description, and a 'See the description of the simulated dataset names in:' section with buttons for 'Dataset', 'Simulated', 'Standard Model Physics', 'Electroweak', and 'CMS'.

Faceted search interface

```
{
  "hits": {
    "hits": [
      {
        "created": "2020-12-21T18:53:13.614856+00:00",
        "sp": 600,
        "url": "https://opendata.cern.ch/api/records/690"
      },
      {
        "metadata": {
          "schema": "http://opendata.cern.ch/schema/records/record-v1.0.0.json",
          "files": {
            "bucket": "84d99807-e73e-6596-8799-2a0f445d99c1",
            "checksum": "a5f83239accac5",
            "key": "8T0u_tg",
            "size": 2678577,
            "version_up": "9d2f67d31-f22a-4e39-8a6f-87d5b3a6372f"
          }
        },
        "abstract": {
          "description": "sgSample event set from /8T0u/Run201808-Apr21Reco-v1/AOD primary dataset in json format readable from the browser-based 3d event display/3d event selection or quality criteria have been applied on the individual physics objects, apart from accepting only the validated runs/sg"
        },
        "accelerator": "CEBN-LHC",
        "authors": [
          {
            "name": "McCauley, Thomas",
            "orcid": "0000-0001-6589-8290"
          }
        ],
        "collections": [
          "CMS-DevProd-003asets"
        ],
        "collision_information": {
          "energy": "7TeV",
          "type": "pp"
        },
        "control_number": "690",
        "date_created": {
          "year": 2020
        },
        "date_published": "2024",
        "distribution": {
          "formats": [
            "tg"
          ],
          "number_events": 25
        },
        "doi": "10.7683/OPENDATA.CMS.6002.0098",
        "experiment": "CMS",
        "files": [
          {
            "bucket": "84d99807-e73e-6596-8799-2a0f445d99c1",
            "checksum": "a5f83239accac5",
            "key": "8T0u_tg",
            "size": 2678577,
            "url": "root://easopds1t.cern.ch/vo/opendata/cms/Run201808/8T0u/16/Apr21Reco-v1/8T0u_tg",
            "version_up": "9d2f67d31-f22a-4e39-8a6f-87d5b3a6372f"
          }
        ],
        "methodology": {
          "description": "These files contain the objects to be displayed with the online event display. No event selection, apart from accepting only the validated runs, is applied. The software to produce these files is available in:",
          "links": [
            {
              "rector": "SG0"
            }
          ]
        }
      }
    ]
  }
}
```

Corresponding REST API

FAIR principles: A is for Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

- ▶ 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

Downloading content

The screenshot shows a web interface for downloading data. A modal window titled "List of files" is open, displaying a table of files with their sizes and download buttons. Below the table is a pagination control with buttons for "1", "2", "3", "4", and "5".

File Name	Size	Action
00E16FB8-9071-E011-83D3-003048673F12.root	583.0 MB	Download
0248915F-EE71-E011-8894-0025902009E8.root	677.0 MB	Download
0268F635-B671-E011-9090-002481E14E00.root	822.5 MB	Download
0278F65A-9A71-E011-A5C0-0025902008A8.root	630.2 MB	Download
02FF3E00-C171-E011-84C7-002590200ADC.root	766.5 MB	Download

Below the modal, the "File Indexes" section shows a list of files with columns for "Filename" and "Size". Each entry includes "List Files" and "Download" buttons. A pagination bar at the bottom shows "First", "Previous", "1", "2", "Next", and "Last".

Browsing and downloading files manually



The cernopendata-client is a command-line tool to interact with the CERN Open Data portal.

Navigation

1. Installation
2. Usage
3. CLI API
4. Changes
5. Contributing
6. License
7. Authors

cernopendata@GitHub
cernopendata@Twitter
opendata-forum.cern.ch
opendata.cern.ch

Quick search

cernopendata-client

python 2.7 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | C | C++ | Go | Java | JavaScript | Kotlin | Lua | Perl | PHP | Python | R | Rust | Scala | Swift | TypeScript | Visual Basic | X++ | YAML

cernopendata-client is a command-line tool to facilitate downloading files from the CERN Open Data portal. The tool enables to query datasets hosted on the CERN Open Data portal and to download and verify the individual data set files.

1. Installation
 - 1.1. PyPI
 - 1.2. Docker
2. Usage
 - 2.1. General help
 - 2.2. Selecting records
 - 2.3. Getting metadata
 - 2.4. Listing available data files
 - 2.5. Downloading data files
 - 2.6. Verifying files
 - 2.7. Listing directories
 - 2.8. More information
3. CLI API
 - 3.1. cernopendata-client
 - 3.1.1. download-files
 - 3.1.2. get-file-locations
 - 3.1.3. get-metadata
 - 3.1.4. list-directory
 - 3.1.5. verify-files
 - 3.1.6. version
4. Changes
 - 4.1. Version 0.2.0 (2020-11-19)
 - 4.2. Version 0.1.0 (2020-09-24)
 - 4.3. Version 0.0.1 (2020-09-09)
5. Contributing
 - 5.1. Issues
 - 5.2. Pull requests
6. License
7. Authors

Automated cernopendata-client

Command-line client

```
/tmp $ 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.

/tmp $ cernopendata-client get-metadata --doi 10.7483/OPENDATA.CMS.ZVT8.MZNY \
- $ --output-value distribution
{
  "formats": [
    "aodsim",
    "root"
  ],
  "number_events": 30125269,
  "number_files": 26958,
  "size": 9560202852603
}

/tmp $ cernopendata-client get-metadata --doi 10.7483/OPENDATA.CMS.ZVT8.MZNY \
--output-value distribution.size
9560202852603
```

Downloading metadata information

```
/tmp $ cernopendata-client get-file-locations --recid 5500
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/BuildFile.xml
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/HiggsDemoAnalyzer.cc
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/List_indexfile.txt
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/M4Lnormdatall.cc
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/M4Lnormdatall_lvl3.cc
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/demoanalyzer_cfg_level3MC.py
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/demoanalyzer_cfg_level3data.py
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/demoanalyzer_cfg_level4MC.py
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/demoanalyzer_cfg_level4data.py
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/mass4l_combine.pdf
http://opendata.cern.ch/eos/opendata/cms/software/HiggsExample20112012/mass4l_combine.png

/tmp $ cernopendata-client download-files --recid 5500 --verify
==> Downloading file 1 of 11
-> File: ./5500/BuildFile.xml
-> Progress: 0/0 KiB (100%)
==> Verifying file BuildFile.xml...
-> Expected size 305, found 305
-> Expected checksum Adler32:ff63668a, found Adler32:ff63668a
==> Downloading file 2 of 11
-> File: ./5500/HiggsDemoAnalyzer.cc
-> Progress: 81/81 KiB (100%)
==> Verifying file HiggsDemoAnalyzer.cc...
-> Expected size 83761, found 83761
-> Expected checksum Adler32:f205f068, found Adler32:f205f068
==> Downloading file 3 of 11
-> File: ./5500/List_indexfile.txt
-> Progress: 1/1 KiB (100%)
==> Verifying file List_indexfile.txt...
-> Expected size 1669, found 1669
-> Expected checksum Adler32:46a907fc, found Adler32:46a907fc
==> Downloading file 4 of 11
-> File: ./5500/M4Lnormdatall.cc
-> Progress: 14/14 KiB (100%)
==> Verifying file M4Lnormdatall.cc...
-> Expected size 14943, found 14943
-> Expected checksum Adler32:af301992, found Adler32:af301992
==> Downloading file 5 of 11
-> File: ./5500/M4Lnormdatall_lvl3.cc
```

Downloading and verifying files

Working with large datasets



opendata CERN Search

Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data

JQCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6/Summer12_DR53X_FU_RD1_STARTS3_V7N-v1/AODSIM, CMS collaboration

Cite as: CMS collaboration (2017). Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.V17S.MCNY

Dataset Simulated Monte Carlo Project CDB CDF ATLAS CMS LHC

Description

Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data. See the description of the simulated dataset names in: [About CMS simulated dataset names](#). These simulated datasets correspond to the collision data collected by the CMS experiment in 2012.

Dataset characteristics

30125209 events, 26958 files, 9.6 TB in total.

System details

Recommended `global tag` for analysis: STARTS3_V27:All
Recommended release for analysis: CMSSW_5_3_32

How were these data generated?

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

Step SIM
Release: CMSSW_5_0_0_patch2
Global Tag: STARTS0_V13:All
Generators: pythia6
Production script ([preview](#))
Generator parameters ([preview](#)) ([link](#))
Output dataset: JQCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6/Summer12-STARTS0_V13-v1/GEN-SIM

Step HLT RECO
Release: CMSSW_5_3_14
Global Tag: STARTS3_V7N:All
Production script ([preview](#))
Configuration file for HLT ([link](#))
Configuration file for RECO ([link](#))
Output dataset: JQCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6/Summer12_DR53X_FU_RD1_STARTS3_V7N-v1/AODSIM

A dataset of 9.6 TB size, 26'958 files

```
/tmp $ cernopendata-client download-files --help
Usage: cernopendata-client download-files [OPTIONS]
```

Download data files belonging to a record.

Select a CERN Open Data bibliographic record by a record ID, a DOI, or a title and download data files belonging to this record.

Examples:

```
$ cernopendata-client download-files --recid 5500
$ cernopendata-client download-files --recid 5500 --filter-name BuildFile.xml
$ cernopendata-client download-files --recid 5500 --filter-regexp py$
$ cernopendata-client download-files --recid 5500 --filter-range 1-4
$ cernopendata-client download-files --recid 5500 --filter-range 1-2,5-7
$ cernopendata-client download-files --recid 5500 --filter-regexp py --filter-range 1-2
```

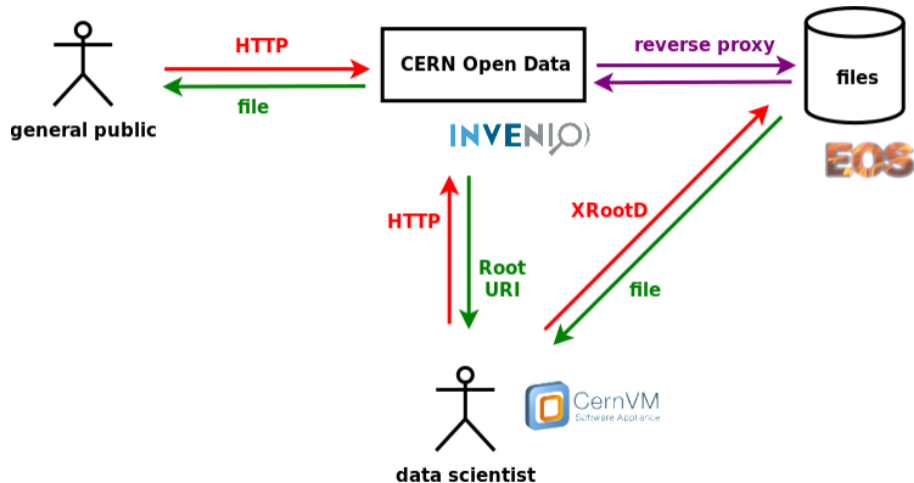
```
/tmp lm 4s $ cernopendata-client get-file-locations --recid 8884 | head -3
http://opendata.cern.ch/eos/opendata/cms/MonteCarlo2012/Summer12_DR53X/QCD_Pt_
v1/00000/0052F9A9-5EB2-E311-81A9-0026189437ED.root
http://opendata.cern.ch/eos/opendata/cms/MonteCarlo2012/Summer12_DR53X/QCD_Pt_
v1/00000/00BF7DF8-69B2-E311-89E6-003048678E8A.root
http://opendata.cern.ch/eos/opendata/cms/MonteCarlo2012/Summer12_DR53X/QCD_Pt_
v1/00000/020194EC-CAB2-E311-8B3A-003048678B94.root
```

Downloading file locations

```
/tmp $ cernopendata-client download-files --recid 8884 --filter-range 1
==> Downloading file 1 of 1
-> File: ./8884/0052F9A9-5EB2-E311-81A9-0026189437ED.root
[ ] -> Progress: 38869/150643 KiB (25%)
```

Downloading one file for inspection

Supporting HTTP and XRootD access protocols



FAIR principles: I is for Interoperable

The data usually need to be integrated with other data.

In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- ▶ 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

Data formats and vocabularies

Filter by file type

<input type="checkbox"/> C	3	<input type="checkbox"/> jpg
<input type="checkbox"/> aod	115	<input type="checkbox"/> json
<input type="checkbox"/> aodsim	849	<input type="checkbox"/> m4v
<input type="checkbox"/> cc	11	<input type="checkbox"/> minlaodsim
<input type="checkbox"/> csv	933	<input type="checkbox"/> nanoaod
<input type="checkbox"/> docx	1	<input type="checkbox"/> ova
<input type="checkbox"/> fevtdebught	1	<input type="checkbox"/> pdf
<input type="checkbox"/> gen-sim	4	<input type="checkbox"/> png
<input type="checkbox"/> gen-sim-dlgl-raw	1	<input type="checkbox"/> py
<input type="checkbox"/> gen-sim-reco	6	<input type="checkbox"/> raw
<input type="checkbox"/> gz	17	<input type="checkbox"/> reco
<input type="checkbox"/> h5	3	<input type="checkbox"/> root
<input type="checkbox"/> html	7	<input type="checkbox"/> tar
<input type="checkbox"/> lg	95	<input type="checkbox"/> tar.gz
<input type="checkbox"/> ipynb	2	<input type="checkbox"/> txt
<input type="checkbox"/> jpg	1	<input type="checkbox"/> xls
<input type="checkbox"/> json	12	<input type="checkbox"/> xml
		<input type="checkbox"/> zip

A variety of data formats
ROOT as a community standard

1	Filter by category	17
12	<input type="checkbox"/> B physics and Quarkonia	19
1	<input checked="" type="checkbox"/> Exotica	18
22	<input type="checkbox"/> Gravitons	1
10	<input type="checkbox"/> Miscellaneous	338
2	<input checked="" type="checkbox"/> Higgs Physics	62
16	<input type="checkbox"/> Beyond Standard Model	276
3	<input type="checkbox"/> Standard Model	2
991	<input type="checkbox"/> Physics Modelling	497
16	<input checked="" type="checkbox"/> Standard Model Physics	39
17	<input type="checkbox"/> Drell-Yan	133
1126	<input type="checkbox"/> ElectroWeak	25
1	<input type="checkbox"/> Forward and Small-x QCD Physics	23
1	<input type="checkbox"/> Minimum Bias	143
28	<input type="checkbox"/> QCD	134
1	<input type="checkbox"/> Top physics	3
6	<input type="checkbox"/> Supersymmetry	
21		

No formal vocabularies
Some common classification

Data semantics

opendata CERN Search Help About

Samples with full event information including tracker hits for tracking, ML, and top quark tagging studies

Usal, Emanuele; Andrews, Michael; Burkle, Bjorn; Gleyzer, Sergei; Narain, Meenakshi

Cite as: Usal, Emanuele; Andrews, Michael; Burkle, Bjorn; Gleyzer, Sergei; Narain, Meenakshi (2019). Samples with full event information including tracker hits for tracking, ML, and top quark tagging studies. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.CHC3.SKPG

Dataset Overview Tracker hit-enriched 300 to 600 bin of QCD_Pt-15to3000_TuneZstar_Flat_8TeV_pythia8
Dataset Overview Tracker hit-enriched 400 to 600 bin of QCD_Pt-15to3000_TuneZstar_Flat_8TeV_pythia8
Dataset Overview Tracker hit-enriched 600 to 3000 bin of QCD_Pt-15to3000_TuneZstar_Flat_8TeV_pythia8

Description

Samples in this record are in a custom root ntuple format and contain the position of the hits and information from the generator-level objects associated to the tracker hits. The samples can be used to study top quark identification algorithms that use low-level detector information such as tracker hits. Machine learning algorithms are suitable for this classification task.

They have been produced from datasets, which consists of events extracted from simulated proton-proton collision events at a center-of-mass energy of 8 TeV generated with Pythia 6 (QCD) or MadGraph2.6 and Pythia6 (top-antitop pair sample). The particles emerging from the collisions traverse through a simulation of the CMS detector.

The parent datasets of these samples contain light jets (QCD) in various energy ranges or all-hadronic high transverse momentum decays of top quarks and consist of hits from the tracking detector, reconstructed tracks, simulated tracks, generated particles, and jets clustered from the generated particles. The various objects are matched in order to reconstruct the provenance of the various hits. Samples are produced from the standard CMS format "AODSIM" plus a series of low-level tracker-related collections that allow the extraction of the tracker hits.

Dataset name	Description	Number of events	Number of files
QCD300to600	QCD, flat pt hat spectrum, 300 < pt hat < 600 GeV	1497600	2496
QCD400to600	QCD, flat pt hat spectrum, 400 < pt hat < 600 GeV	1989000	3315
QCD600to3000	QCD, flat pt hat spectrum, 600 < pt hat < 3000 GeV	2974800	4959
tbar	tbar, fully hadronic decays, pt of the top/antitop greater than 400 GeV	2969109	4055

Related datasets

QCD300to600_Run1_8TeV was derived from:
Tracker-hit-enriched 300 to 600 bin of QCD_Pt-15to3000_TuneZstar_Flat_8TeV_pythia8

QCD400to600_Run1_8TeV was derived from:
Tracker-hit-enriched 400 to 600 bin of QCD_Pt-15to3000_TuneZstar_Flat_8TeV_pythia8

QCD600to3000_Run1_8TeV was derived from:
Tracker-hit-enriched 600 to 3000 bin of QCD_Pt-15to3000_TuneZstar_Flat_8TeV_pythia8

tbar_Run1_8TeV was derived from:
Tracker-hit-enriched TJets_HadronicMGDecays_8TeV-madgraph

Dataset characteristics

9430509 events, 14825 files, 12.8 TB in total.

Dataset semantics

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 TOR, 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?
hit_genparticle_id	std::vector<unsigned int>	Index of the gen particle matched to the hit
hit_pdgid	std::vector<int>	PDG ID of the gen particle matched to the hit
hit_recotrack_id	std::vector<unsigned int>	Index of the reco track matched to the hit
hit_recotrack_match	std::vector<bool>	is the hit matched to a reco track?
hit_genparticle_match	std::vector<bool>	is the hit matched to a gen particle?
hit_genjet_id	std::vector<unsigned int>	Index of the gen jet matched to the hit

Dataset variables coming with detailed semantics description

Interoperability via accompanying examples

opendata CERN Search Help About

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

Wunsch, Stefan

Cite as: Wunsch, Stefan; (2019). Analysis of the di-muon spectrum using data from the CMS detector taken in 2012. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.AAR1.AN2Q

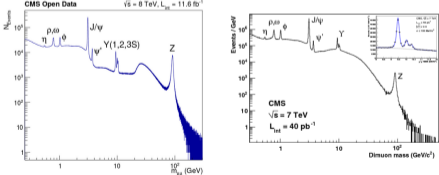
Selection Analysis Results Data

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 J/ψ 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:



Dimuon spectrum analysis example

opendata CERN Search Help About

Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012

Wunsch, Stefan

Cite as: Wunsch, Stefan; (2019). Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.GV20.PR5T

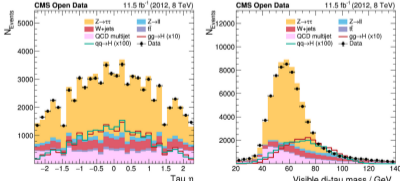
Selection Analysis Results Data

Description

This analysis uses data and simulation of events at the CMS experiment from 2012 with the goal to study decays of a Higgs boson into two tau leptons in the final state of a muon lepton and a hadronically decayed tau lepton. The analysis follows loosely the setup of the official CMS analysis published in 2014.

The purpose of the original CMS analysis was to establish the existence of the Higgs boson decaying into two tau leptons. Since performing this analysis properly with full consideration of all systematic uncertainties is an enormously complex task, we reduce this analysis to the qualitative study of the kinematics and properties of such events without a statistical analysis. However, as you can explore in this record, already such a reduced analysis is complex and requires extensive physics knowledge, which makes this a perfect first look into the procedures required to claim the evidence or existence of a new particle.

Two example results produced by this analysis can be seen below. The plots show the data recorded by the detector compared to the estimation of the contributing processes, which are explained in the following. The analysis has implemented the visualization of 34 such observables.



$H \rightarrow \tau\tau$ analysis example

FAIR principles: R is for Reusable

The ultimate goal of FAIR is to optimise the reuse of data.

To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- ▶ 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

Data provenance: how were the data generated?

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

/BulkGravTohhTohbbbbb_narrow_M-4500_13TeV-madgraph/RunII/Summer16/MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TracheIV_v6-v1/MINIAODSIM, CMS Collaboration

Cite as: CMS Collaboration (2019). Simulated dataset BulkGravTohhTohbbbbb_narrow_M-4500_13TeV-madgraph in MINIAODSIM format for 2016 collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.7N4X.Z7FA

[Dataset](#) [Simulated](#) [Exotica](#) [Gravitons](#) [CMS](#) [13TeV](#) [CERN-LHC](#)

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: /BulkGravTohhTohbbbbb_narrow_M-4500_13TeV-madgraph/RunII/Winter15wmLHE-MCRUN2_71_V1-v1/LHE

Note: To get the exact generator parameters, please see [Finding the generator parameters](#).

Step SIM

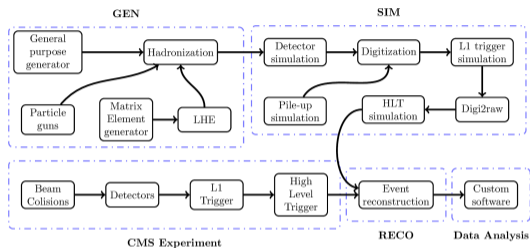
Release: CMSSW_7_1_20

Configuration file for SIM ([link](#))

Output dataset: /BulkGravTohhTohbbbbb_narrow_M-4500_13TeV-madgraph/RunII/Summer15GS-MCRUN2_71_V1-v1/GEN-SIM

Step HLT RECO

Release: CMSSW_8_0_71



- ▶ full capture of data generation steps
- ▶ full capture of compute environments
- ▶ full capture of configuration files
- ▶ full capture of production scripts

Data records come with full provenance information

Data provenance is part of the standard JSON metadata

```
"methodology": {
  "description": "<p>These data were generated in several steps (see also <a href=\"/docs/cms-mc-production-overview\">CMS M
  "steps": [
    {
      "configuration_files": [
        {
          "script": "#!/bin/bash\nsource /cvmfs/cms.cern.ch/cmsset_default.sh\nexport SCRAM_ARCH=slc5_amd64_gcc462\nif [ -r
          "title": "Production script"
        },
        {
          "title": "Generator parameters",
          "url": "https://cms-pdmv.cern.ch/mcm/public/restapi/requests/get\_fragment/HIG-Summer12-02276"
        },
        {
          "cms_confdb_id": "a97a2f6c22dfba999c0131657a81ecfd",
          "process": "SIM",
          "title": "Configuration file"
        }
      ],
      "generators": [
        "pythia6"
      ],
      "global_tag": "START53_V7C::All",
      "output_dataset": "/BBH_HToTauTau_M_125_TuneZ2star_8TeV_pythia6_tauola/Summer12-START53_V7C-v1/GEN-SIM",
      "release": "CMSSW_5_3_13",
      "type": "SIM"
    },
    {
      "configuration_files": [
```

CERN Open Data portal standardises CMS DAS/McM information

Can we reprocess published data samples?

SingleElectron primary dataset sample in RAW format from RunA of 2011 (from /SingleElectron/Run2011A-v1/RAW)

/SingleElectron/Run2011A-v1/RAW, CMS collaboration

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.6O84.WLN8

Dataset Collision CMS 7TeV CERN-LHC

Description

A sample from SingleElectron primary dataset in RAW format from RunA of 2011. Run range [161224,163286].

This dataset contains selected runs from 2011 RunA. The list of validated lumi sections, which must be applied to all analyses on events reconstructed from these data, can be found in

[CMS list of validated runs Cert_160404-180252_7TeV_ReRecoNov08_Collisions11_JSON.txt](#)

Dataset characteristics

2064298 events. **116** files. **424.3 GB** in total.

How can you use these data?

These data are in RAW format and not directly usable in analysis. The reconstructed data reprocessed from these RAW data are included in the data of [this record](#). The reconstruction step can be repeated with the configuration file below and the resulting AOD has been confirmed to be identical with the original one with comparison code available in [Validation code to plot basic physics objects from AOD](#)

RAW

SingleElectron primary dataset in AOD format from RunA of 2011 (/SingleElectron/Run2011A-12Oct2013-v1/AOD)

/SingleElectron/Run2011A-12Oct2013-v1/AOD, CMS collaboration

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 Collision CMS 7TeV CERN-LHC

Description

SingleElectron primary dataset in AOD format from RunA of 2011. Run period from run number 160404 to 173692.

This dataset contains all runs from 2011 RunA. The list of validated runs, which must be applied to all analyses, can be found in

[CMS list of validated runs Cert_160404-180252_7TeV_ReRecoNov08_Collisions11_JSON.txt](#)

Dataset characteristics

41709195 events. **1542** files. **5.8 TB** in total.

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](#).

Data processing / RECO

This primary AOD dataset was processed from the RAW dataset by the following step:
Step: RECO

Release: CMSSW_5_3_12_patch1

Global tag: FT_R_53_LVS:All

[Configuration file for RECO step reco_2011A_SingleElectron](#)

AOD

Reprocessing environment: containers

The screenshot shows the Docker Hub page for the repository `atlas/analysisbase`. The repository is owned by `atlas` and was last updated 4 days ago. It contains four tags: `21.6.184`, `21.6.183`, `21.6.182`, and `21.6.181`. Each tag has a corresponding Dockerfile and a size of 806.8 MB. The repository is categorized under 'Container'.

ATLAS collaboration

<https://hub.docker.com/r/atlas/analysisbase/tags>

The screenshot shows the Docker Hub page for the repository `cms-sw/docker`. The repository is owned by `cms-sw` and was last updated 10 days ago. It contains four tags: `9.2.3`, `9.2.2`, `9.2.1`, and `9.2.0`. Each tag has a corresponding Dockerfile and a size of 806.8 MB. The repository is categorized under 'Container'.

CMS collaboration

<https://gitlab.cern.ch/cms-cloud/cmssw-docker>

Docker and Singularity images help to encapsulate the computing environment

Reprocessing workflow: computational recipe

3. Workflow

The workflow can be logically divided into several parts:

0. Upload all files.

Some files cannot be generated at run time and need to be uploaded.

```
inputs:
files:
- src/PhysicsObjectsHistos.cc
- BuildFile.xml
- demoanalyzer_cfg.py
```

1. Fix the CMS SW environment variables manually.

First, we have to set up the environment variables accordingly for the [CMS SW](#). Although this is done in the docker image, reana overrides them and they need to be reset. This is done by invoking the `cms entrypoint.sh` script commands.

See also [this issue](#).

```
$ source /opt/cms/cmsset_default.sh
$ scramv1 project CMSSW CMSSW_5_3_32
$ cd CMSSW_5_3_32/src
$ eval `scramv1 runtime -sh`
```

2. Create the specific CMS path.

CMS specific data analysis framework requires two directory levels. See also [this issue](#).

```
$ mkdir Reconstruction && cd Reconstruction
$ mkdir Validation && cd Validation
```

3. Create the reconstruction file.

See also [this repo](#).

```
$ cmsDriver.py reco -s RAM2DIGI,L1Reco,RECO,USER:EventFilter/HcalRawToDigi/hcallaserbhbheffilter2012_cf
```

4. Adjust the reconstruction file to the specific data file.

Although generated using parameters, the reconstruction file still requires changes.

```
$ sed -i 's/from Configuration.AiCa.GlobalTag import GlobalTag/process.GlobalTag.connect = cms.string("
$ sed -i 's/# Other statements/from Configuration.AiCa.GlobalTag import GlobalTag/g' reco_cmsdriver.py
$ sed -i 's/process.GlobalTag = GlobalTag(process.GlobalTag, 'FT_53_LV5_ANI::All', '')/process.GlobalTag
```

5. Link the CVMFS files.

The `-l` commands are explicitly needed to make sure that the `cms-opendata-conddb.cern.ch` directory has actually expanded in the image, according to [this guide](#). See also [this issue](#).

```
$ ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_ANI_RUNA_FT_53_LV5_ANI
$ ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_ANI_RUNA.db_FT_53_LV5_ANI_RUNA.db
$ ls -l
$ ls -l /cvmfs/
```

6. Run the reconstruction.

At this point all environment variables and files should be proper.

```
$ cmsRun reco_cmsdriver.py
```

7. Adjust project structure for validation

Copy the required files for the next steps.

```
$ mkdir src
$ scp ../../../../src/PhysicsObjectsHistos.cc ./src
$ scp ../../../../BuildFile.xml .
$ scp ../../../../demoanalyzer_cfg.py .
```

8. Run CMS scram command to fix libraries.

Most importantly, the `BuildFile.xml` has to be inside the directory where the `scram` command is executed.

```
$ scram b
```

9. Run the validation file.

See also [this repo](#)

```
$ cmsRun demoanalyzer_cfg.py
```

Workflow steps to run CMS reconstruction in CMSSW environment

Four pillars of reusable computational research

I. Input data

What is your input data?

- input files
- input parameters

II. Analysis code

Which code analyses it?

- user code
- software frameworks

III. Computing environment

What is your environment?

- operating system
- database calls

IV. Computational recipes

Which steps did you take?

- shell commands
- notebooks and workflows

REANA: A platform for reproducible analyses

reana

[Home](#) [Examples](#) [Get Started](#) [Documentation](#) [News](#) [Contact](#)

reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.



Scalable

Support for remote compute clouds.



Reusable

Containerise once, reuse elsewhere. Cloud-native.



Free

Free Software. MIT licence. Made with ❤️ at CERN.



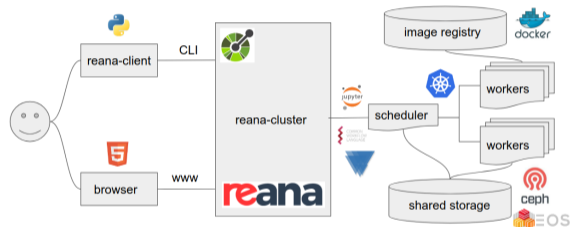
<http://www.reana.io>

Install and run containerised workflows on Kubernetes, HTCondor, Slurm backends

REANA: Running containerised workflows on remote compute clouds

```
1 version: 0.6.0
2 inputs:
3 files:
4 - skin.cxx
5 - histograms.py
6 - plot.py
7 workflow:
8 type: serial
9 specification:
10 steps:
11 - name: compiling
12   environment: reanahub/reana-env-root6:6.18.04
13   commands:
14     - g++ -g -O3 -Wall -Wextra -Wpedantic -o skin skin.cxx `root-config --cflags --libs`
15 - name: skimming
16   environment: reanahub/reana-env-root6:6.18.04
17   commands:
18     - ./skin
19 - name: histogramming
20   environment: reanahub/reana-env-root6:6.18.04
21   commands:
22     - python histograms.py
23 - name: plotting
24   environment: reanahub/reana-env-root6:6.18.04
25   commands:
26     - python plot.py
```

<https://github.com/cms-opendata-analyses>

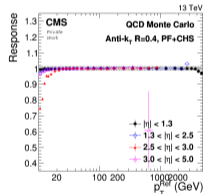
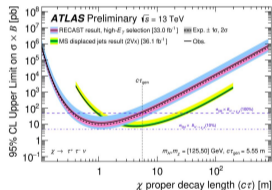
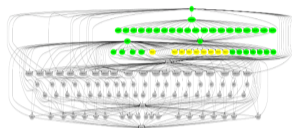


<https://doi.org/10.1051/epjconf/201921406034>

Structure data analysis by means of a
reana.yaml file

Use command-line and web interfaces to
launch analysis on remote compute clusters

Computational recipes: notebooks & workflows



ATLAS <http://cdsweb.cern.ch/record/2714064>

CMS <https://github.com/alintulu/reana-demo-JetMETAnalysis>

Data analysis example: ATLAS displaced jet search reinterpretation

Data production example: CMS jet energy resolution and corrections

An ecosystem of sister data repositories

HEPData widens publication-level data scope

HEPData Search HEPData

Reconstruction and identification of boosted di- τ systems in a search for Higgs boson pairs using 13 TeV proton-proton collision data in ATLAS

Table 1

Stage of selection	Pre-selection	Di- τ selection	Large-JT jet selection	Signal region
1000	0.245581	0.0517972	0.0122319	0.0020
1100	0.033949	0.070880	0.0320259	0.0089
1200	0.403440	0.162247	0.0550608	0.0183
1400	0.408088	0.130511	0.0090454	0.0155
2000	0.059782	0.139030	0.11138	0.0150

ATLAS collaboration <https://www.hepdata.net/record/ins1809175>

HEPData provides interactive interface to explore and download publication-level data behind plots and tables

HEPData Search HEPData

Search for bottom-squark pair production in the ATLAS detector in final states containing Higgs bosons, b-jets and missing transverse momentum

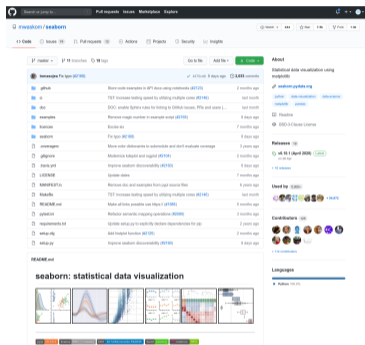
Table 1

cmenergies	pframes	reactions
2000	SCF	PP -> BOTTOM BOTTOM

ATLAS collaboration <https://www.hepdata.net/record/ins1748602>

... now starting to handle more data types: likelihoods!

Zenodo preserves research software



Latest release

v0.10.1
dd40fd6

DOI: 10.5281/zenodo.3767070



<https://guides.github.com/activities/citable-code>

GitHub ↔ Zenodo bridge to automatically preserve releases

Preserving “restricted” analysis knowledge



Home What is CAP? Get Started Integrations Documentation Log In

CERN Analysis Preservation

capture, preserve and reuse physics analyses



Capture



Collect and preserve elements
needed to understand and
rerun your analyses

Collaborate



Share your analysis and
components with other users,
your collaboration or group

Reuse



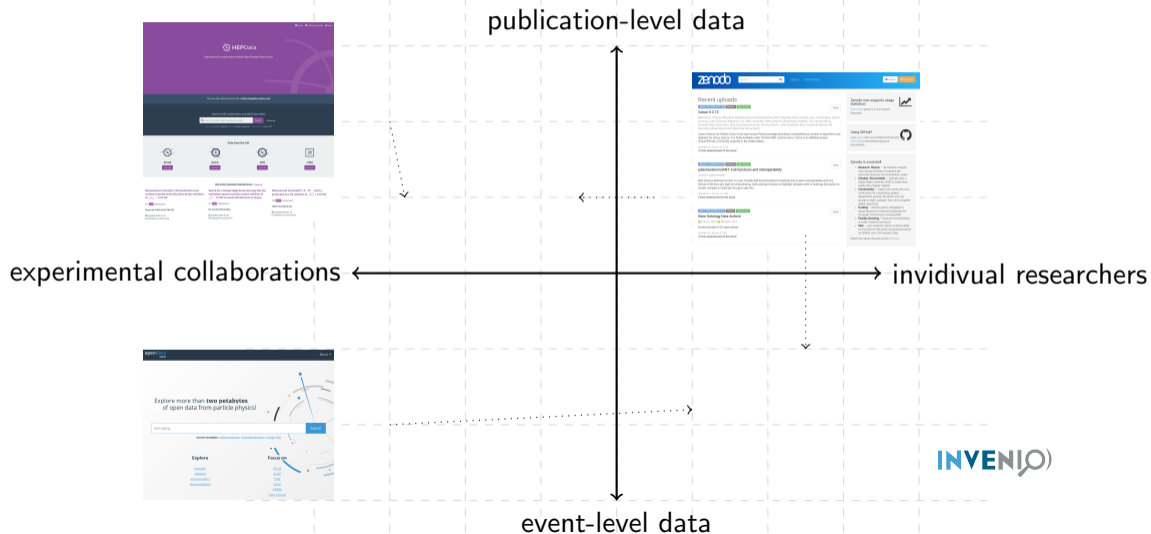
Run containerized workflows
and easily reuse analysis
components

The screenshot shows the CERN Analysis Preservation web interface. It features a navigation bar at the top with links for Home, What is CAP?, Get Started, Integrations, Documentation, and Log In. The main content area is divided into two columns: Filters and Results. The Filters column includes sections for TYPE, COLLISION SYSTEM, CADI STATUS, and ACCELERATOR PARAMETERS, each with a list of checkboxes and their corresponding labels. The Results column displays a list of analysis entries, each with a title, date, and a small icon. Below the Results section, there is a terminal window showing a series of commands and their outputs, including the installation of the csp-client and the successful upload of a file.

<https://analysispreservation.cern.ch>

CERN Analysis Preservation framework for collaboration-restricted data.
Following the same FAIR principles (FAIR \neq open!)

A family of digital repositories in movement



Conclusions

Fostering FAIR science



CERN Open Data



CERN Analysis Preservation



REANA Reusable Analyses

“Capturing and sharing actionable data and knowledge behind particle physics data analyses in order to facilitate future data reuse”

CERN IT A. Mečionis, D. Rodríguez, P. Shandilya, T. Šimko, M. Vidal García · **CERN SIS** P. Fokianos, I. Koutsakis, K. Naim, A. Papadopoulos · **ALICE** Y. Foka, M. Gheata, C. Grigoras, M. Zimmermann · **ATLAS** K. Cranmer, L. Heinrich, A. Sanchez Pineda, D. Rousseau, F. Socher · **CMS** H. Bittencourt, A. Calderon, E. Carrera, A. Geiser, A. Huffman, C. Lange, K. Lassila-Perini, L. Lloret, T. McCauley, A. Rao, A. Rodriguez Marrero · **LHCb** S. Amerio, C. Burr, B. Couturier, S. Neubert, C. Parkes, S. Roiser, A. Trisovic · **OPERA** G. De Lellis, S. Dmitrievsky, G. Galati, A. Ustyuzhanin · **HepData** A. Clarke, E. Maguire, G. Watt · **Invenio** L. H. Nielsen, P. Panero · **Zenodo** F. Decourcelle, A. Ioannidis, G. Lignos · **CERN CernVM** J. Blomer · **CERN EOS** L. Mascetti, H. Rousseau · **CERN Kubernetes** R. Rocha · **CERN OpenShift** A. Lossent

Free open-source solutions

The screenshot shows the GitHub repository page for 'CERN Open Data'. The repository is located at `https://github.com/cernopendata/`. It features a dark header with navigation links for Pull requests, Issues, Marketplace, and Explore. The repository name 'CERN Open Data' is prominently displayed, along with a description: 'Explore more than two petabytes of open data from particle physics!'. The location is listed as Geneva, Switzerland, and the website is `http://opendata.cern.ch`. Below the repository name, there are tabs for Repositories (30), Packages, People (50), Teams (3), Projects (4), and Settings. The 'Pinned repositories' section shows two pinned repositories: 'opendata.cern.ch' (Python, 407 stars, 124 forks) and 'data-curation' (Python, 11 stars, 1 fork). A search bar for repositories is present, along with filters for Type and Language. The repository details for 'opendata.cern.ch' are shown, including a list of tags like 'python', 'task', 'big-data', 'json-schema', 'open-data', 'open-science', and 'research-data-management'. It also shows the license as GPL 2.0, 124 stars, 137 forks, and was updated 2 days ago. A 'Top languages' section shows Python, Shell, and JavaScript. A 'People' section shows 50 contributors.

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

CERN Open Data on GitHub

The screenshot shows the CERN Open Data user forum. The forum is located at `https://opendata-forum.cern.ch`. It features a blue header with the 'opendata' logo and navigation links for Sign Up and Log In. The forum is organized into categories, with 'Latest' selected. The forum posts are listed in a table with columns for Topic, Replies, Views, and Activity. The posts include:

- Welcome to the CERN Open Data forum! (0 replies, 239 views, Dec '19)
- Error processing specific trigger analysis (3 replies, 107 views, Oct '20)
- X11 forwarding with docker (1 reply, 244 views, Sep '20)
- Error response from daemon: pull access denied for 3813b5241687, repository does not exist or may require 'docker login' (5 replies, 121 views, Sep '20)
- Git usage in CMS open data environments (1 reply, 160 views, Jul '20)
- CMS open data VM images updated (0 replies, 174 views, Jul '20)
- Running CMS OpenData containers in WSL2 (0 replies, 243 views, Jun '20)

<https://opendata-forum.cern.ch>

CERN Open Data user forum

CERN Open Data portal: future plans

- ▶ December 2020: A common statement on the open data policy by CERN management and ATLAS, ALICE, CMS, LHCb and TOTEM experiments.

<https://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments>

- ▶ Prepare for forthcoming increase in open data publishing.
- ▶ Introduce flexible hot/cold disk/tape storage solution. Part of dataset files on disk, part on tapes.
- ▶ Simplify ingestion and exposure of experiment datasets (Rucio, Dirac).
- ▶ Further automatise provenance testing and usage examples.

CERN announces new open data policy in support of open science

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

11 DECEMBER, 2020



Data storage solutions at the CERN data centre (Image: CERN)

Geneva, 11 December 2020. The four main LHC collaborations (ALICE, ATLAS, CMS and LHCb) have unanimously endorsed a new open data policy for scientific experiments at the Large Hadron Collider (LHC), which was presented to the CERN Council today. The policy commits to publicly releasing so-called level 3 scientific data, the type required to make scientific studies, collected by the LHC experiments. Data will start to be released approximately five years after collection, and the aim is for the full dataset to be publicly available by the close of the experiment concerned. The policy addresses the growing movement of open science, which aims to make scientific research more reproducible, accessible, and collaborative.

The level 3 data released can contribute to scientific research in particle physics, as well as research in the field of scientific computing, for example to improve reconstruction or analysis methods based on machine learning techniques, an approach that requires rich data sets for training and validation.