

Automatising open data publishing workflows: experience with CMS open data curation

Kati Lassila-Perini¹ Tibor Šimko²

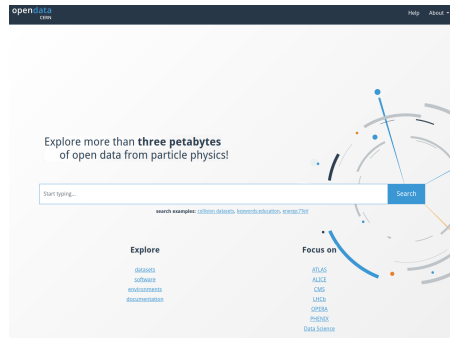
on behalf of the CMS Collaboration and the CERN Open Data team

¹HIP ²CERN

*26th International Conference on Computing in High Energy and Nuclear Physics (CHEP)
Norfolk, United States, 8–12 May 2023*

CERN Open Data portal

- ▶ digital repository for event-level particle physics open data
 - collision and simulated datasets for research
 - derived datasets for education
 - configuration files and documentation
 - virtual machines and container images
 - software tools and analysis examples
- ▶ launched in November 2014
- ▶ total size in April 2023
 - over 15 000 bibliographic records
 - over 1 500 000 files
 - over 3 petabytes

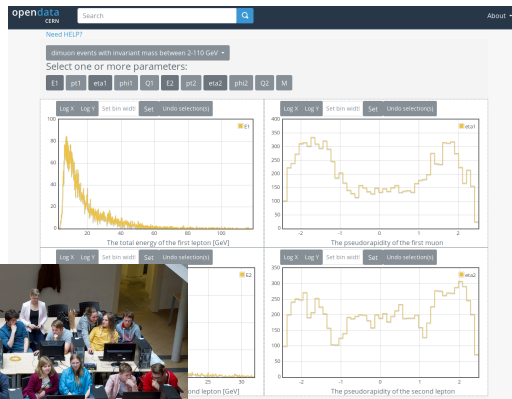
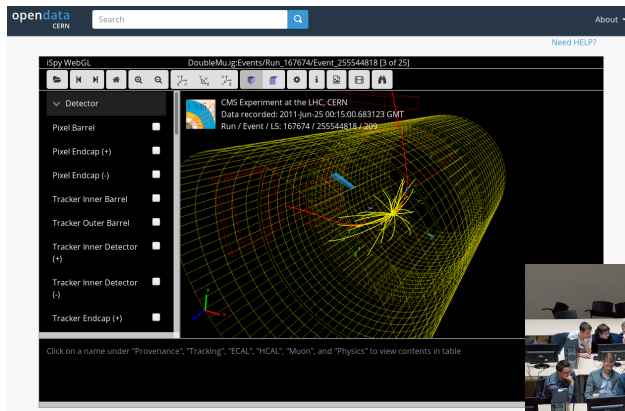


<https://opendata.cern.ch>

Developed by CERN in close collaboration with Experiments



Education-oriented use cases



Interactive event display and histogramming for derived datasets

Research-oriented use cases

docker hub

Explore cmsopendata/cmssw_7_6_7-slc6_amd64_gcc493

cmsopendata/cmssw_7_6_7-slc6_amd64_gcc493

By CMS Collaboration • Updated a month ago

Overview Tags

Sort by Newest Filter Tags

```
ls -l /cvmfs/cms.opendata.cern.ch/
total 1855262
drwxr-xr-x. 2 cvmfs cvmfs 4096 Jan 21 2016 FT_53_LVS_AHL
drwxr-xr-x. 2 cvmfs cvmfs 4096 Feb 22 2016 FT_53_LVS_AHL_RUNA
drwxr-xr-x. 2 cvmfs cvmfs 4096 Jun 23 2017 FT53_V21A_AHL
drwxr-xr-x. 2 cvmfs cvmfs 4096 Nov 29 2017 FT53_V21A_AHL_FULL
drwxr-xr-x. 2 cvmfs cvmfs 4096 Jun 23 2017 FT53_V21A_AHL_RUNC
drwxr-xr-x. 2 cvmfs cvmfs 4096 Oct 20 2017 FT_R_42_V18A
drwxr-xr-x. 2 cvmfs cvmfs 4096 Nov 9 2018 START42_V21B
drwxr-xr-x. 2 cvmfs cvmfs 4096 Jun 23 2016 START53_V16A1
drwxr-xr-x. 2 cvmfs cvmfs 4096 Jun 23 2017 START53_V27
drwxr-xr-x. 2 cvmfs cvmfs 4096 Nov 30 2018 START53_V78
drwxr-xr-x. 1 cvmfs cvmfs 1082414088 Oct 31 2018 1025_upgrade2018_design_v9.db
drwxr-xr-x. 1 cvmfs cvmfs 691543216 Oct 31 2018 BOX_sclun2_asymptotic_2016_TracheIV_v8.db
drwxr-xr-x. 1 cvmfs cvmfs 82944 Jan 21 2016 FT_53_LVS_AHL.db
drwxr-xr-x. 1 cvmfs cvmfs 82944 Feb 22 2016 FT_53_LVS_AHL_RUNA.db
drwxr-xr-x. 1 cvmfs cvmfs 118088 Jun 23 2017 FT53_V21A_AHL.db
drwxr-xr-x. 1 cvmfs cvmfs 120632 Nov 29 2017 FT53_V21A_AHL_FULL.db
drwxr-xr-x. 1 cvmfs cvmfs 120632 Jun 23 2017 FT53_V21A_AHL_RUNC.db
drwxr-xr-x. 1 cvmfs cvmfs 64512 Oct 20 2017 FT_R_42_V18A.db
drwxr-xr-x. 1 cvmfs cvmfs 72704 Nov 9 2018 START42_V21B.db
drwxr-xr-x. 1 cvmfs cvmfs 84992 Jun 23 2016 START53_V16A1.db
drwxr-xr-x. 1 cvmfs cvmfs 130048 Jun 23 2017 START53_V27.db
drwxr-xr-x. 1 cvmfs cvmfs 89088 Nov 30 2018 START53_V78.db
```

cmsopendata/cmssw_7_6_7-slc6_amd64_gcc493

COMPRESSED SIZE 6.58 GB

Use containerised CMSSW environments
with CVMFS condition data snapshots

opendata CERN

Search

About

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

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

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JKB8.RR42

Software Analysis CMS Accelerator CERN.LHC

Description

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

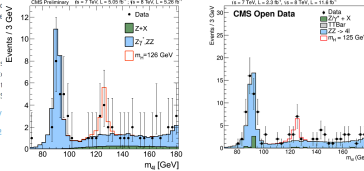
The published reference plot which is being approximated in this example is https://inspirehep.net/record/1124338/files/H4l_mass_3.png. Other Higgs final states (e.g. Higgs to two photons), which were also part of the same CMS paper and strongly contributed to the Higgs boson discovery, are not covered 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 exercises with the linux ops

Use with

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

/DoubleElectron/
/DoubleMu/Run2



icquaintance

re original
again close to,
y as they are,

Study physics analysis examples

Enables independent research

CMS Open Data Workshop 2022
CERN
Aug 1-4, 2022
2:30 pm - 6:30 pm (CET)

Instructors: M. Bellis, E. Carrera, A. Geiser, J. Hogan, C. Lange, K. Lassila-Perini, T. McCauley, S. Sekmen, X. Tintin, J. Yoo

Helpers: A. Chicaiza, K. Chicaiza, N. Dhingra, E. Jimenez, K. Johnson, D. Li, P. Lucas, S. Matham, D. Mene, D.

Monday


14:30-14:50	Welcome and Intro	K. Lassila-Perini
14:50-15:50	Physics Objects: Intro and POET	M. Bellis, E. Carrera, K. Lassila-Perini
15:50-16:30	Physics Objects: Electrons	M. Bellis, E. Carrera, K. Lassila-Perini
16:30-17:00	Break	
17:00-17:40	Physics Objects: Muons	M. Bellis, E. Carrera, K. Lassila-Perini
17:40-18:30	Physics Objects: Jets	M. Bellis, E. Carrera, K. Lassila-Perini

Tuesday

14:30-15:30	Trigger	E. Carrera
15:30-16:30	Luminosity	J. Yoo
16:30-17:30	Break	
17:00-18:30	Analysis example with Run 1 data	M. Bellis, A. Geiser

Wednesday

14:30-15:15	Simplified Run 2 analysis: 1	C. Lange, K. Lassila-Perini, X. Tintin
15:10-16:30	Simplified Run 2 analysis: 2	
16:30-17:00	Break	
17:30-18:30	Simplified Run 2 analysis: 3	S. Sekmen



INSPIRE HEP

70 results | [cite all](#) | [Create Summary](#) | [Most Cited](#)

Fast and Accurate Simulation of Particle Detectors Using Generative Adversarial Networks
Pasquale Musella (ETH, Zurich (main)), Francesco Pandolfi (INFN, Rome) (May 2, 2020)
Published in: Comput.Softw.Big Sci. 2 (2018) 1, 8 • e-Print: 1905.00950 [hep-ph]
[pdf](#) [links](#) [DOI](#) [cite](#) [clean](#) [reference search](#) [66 citations](#)

Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2
LHC Reinterpretation Forum Collaboration • Waleed Abdallah (Harish-Chandra Res. Inst. and Cairo U.) et al. (Mar 15, 2020)
Published in: SciPost Phys. 9 (2020) 2, 022 • e-Print: 2003.07868 [hep-ph]
[pdf](#) [links](#) [DOI](#) [cite](#) [clean](#) [reference search](#) [59 citations](#)

Exposing the QCD Splitting Function with CMS Open Data
Andrew Larkoski (Johannesburg), Simone Morzanti (SUNY, Buffalo), Jesse Thaler (MIT, Cambridge, CTP), Aadish Tripathi (MIT, Cambridge, CTP), Wei Xue (MIT, Cambridge, CTP) (Apr 17, 2017)
Published in: Phys.Rev.Lett. 119 (2017) 13, 132003 • e-Print: 1704.05066 [hep-ph]
[pdf](#) [links](#) [DOI](#) [cite](#) [clean](#) [reference search](#) [58 citations](#)

Interaction networks for the identification of boosted $H \rightarrow b\bar{b}$ decays
Erik A. Moreno (Caltech), Thong Q. Nguyen (Caltech), Jean-Roch Vignati (Caltech), Orlin Cseri (Caltech), Harvey B. Newman (Caltech) et al. (Sep 26, 2019)
Published in: Phys.Rev.D 102 (2020) 1, 012010 • e-Print: 1909.12289 [hep-ph]
[pdf](#) [links](#) [DOI](#) [cite](#) [clean](#) [reference search](#) [48 citations](#)

Unveiling hidden physics at the LHC
Oliver Fischer (Liverpool U.), Bruce Mellado (U. Wolleraub), Johannesburg, Sch. Phys. and (Theoria LABS), Stefan Antusch (Bosch U.), Emanuele Bagdasarian (PSI, Villigen), Shrikha Banerjee (CERN) et al. (Sep 13, 2021)
Published in: Eur.Phys.J.C 62 (2022) 8, 665 • e-Print: 2209.08265 [hep-ph]
[pdf](#) [links](#) [DOI](#) [cite](#) [clean](#) [reference search](#) [38 citations](#)

Exploring the Space of Jets with CMS Open Data
Patrick T. Komiske (MIT, Cambridge, CTP and Harvard U.), Radha Mandlekar (MIT, Cambridge, CTP), Eric M. Metodiev (MIT, Cambridge, CTP and Harvard U.), Priyanka Nair (MIT, Cambridge, CTP), Jesse Thaler (MIT, Cambridge, CTP and Harvard U.) (Aug 22, 2019)
[pdf](#) [links](#) [DOI](#) [cite](#) [clean](#) [reference search](#) [30 citations](#)

Document Type

<input type="checkbox"/> article	48
<input type="checkbox"/> published	26
<input type="checkbox"/> conference paper	22
<input type="checkbox"/> thesis	3
<input type="checkbox"/> review	3

Author

<input type="checkbox"/> Jesse Thaler	11
<input type="checkbox"/> Kati Lassila-Perini	9
<input type="checkbox"/> Clemens Gregor Lange	6
<input type="checkbox"/> L. Lloret Iglesias	4
<input type="checkbox"/> Lukas A. Heinrich	4
<input type="checkbox"/> Patrick T. Komiske III	4

CMS open data workshops for research use

Over seventy papers citing CMS open data

What does it take to make a new CMS open data release?

1. Prepare the release approval within the experiment, evaluate luminosity and data volume
2. Transfer data to be released from CMS storage to EOS open data storage
 - datasets themselves
 - dynamic data: condition database snapshots
 - additional data assets: luminosity information, list of validated runs
3. Collect and prepare metadata
 - content: author, title, number of events, file sizes, etc
 - provenance: how were these data selected? HLT, RECO, configurations, etc
 - usage in a research context: global tag, CMSSW version, luminosity, corrections, etc
4. Prepare the compute environment
 - container images and virtual machines
5. Prepare and test data usage instructions
 - getting started instructions
 - software and workflow examples
6. Mint DOIs and release to public

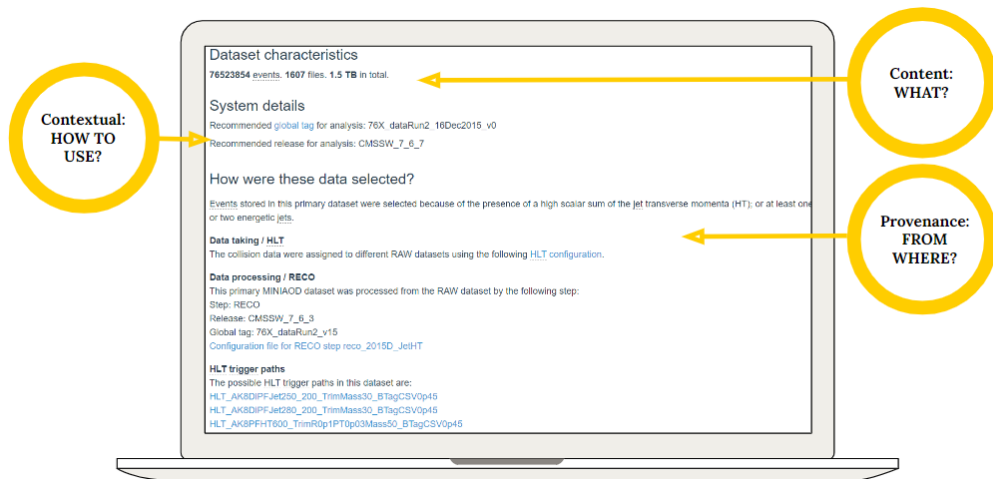
Part I: Data management

- ▶ data files
 - using EOS open data storage volumes ensuring scalability
 - using Rucio T3 endpoint facilitating data transfers in view of open data publishing
- ▶ “dynamic” data
 - need to capture condition db as sqlite snapshots
- ▶ additional data assets
 - luminosity information
 - certified data filters
 - scale and correction factors

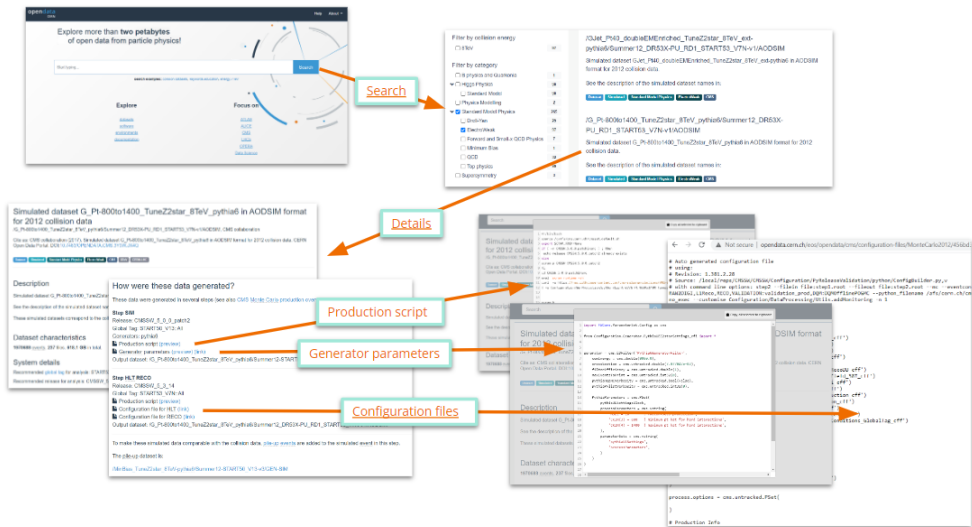


```
/eos/opendata/cms/Run2010A
/eos/opendata/cms/Run2010B
...
/eos/opendata/cms/MonteCarlo2010
...
/eos/opendata/cms/configuration-files
...
/eos/opendata/cms/lhe_generators
...
/eos/opendata/cms/conddb
...
```

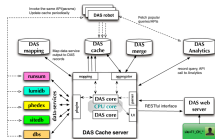

Metadata types



Example: Data provenance of simulated datasets



Capturing data provenance via ad-hoc curation scripts



CMS DAS



CMS McM

```
128 def get_prepId_from_das(dataset, das_dir):
129     "get prepId for dataset"
130
131     # get prepId from das/dataset
132     prepId = get_from_deep_json(get_das_store_json(dataset, 'dataset', das_dir), 'prep_id')
133
134     if prepId == None:
135         # try to get from das/mcm:
136         prepId = get_from_deep_json(get_das_store_json(dataset, 'mcm', das_dir), 'prepId')
137         # todo also try different queries from the json. prep_id?
138
139     return prepId
```

Mining several CMS collaboration sources

Challenges in collecting provenance meta-data

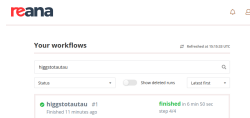
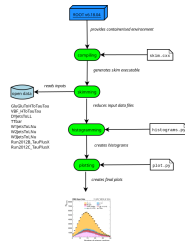
- ▶ in CMS information systems, metadata is stored (or aggregated) by the dataset
- ▶ we need information about full processing chain
- ▶ need to query information through parent relations
- ▶ need to adapt to the evolution of the processing chain and tools

Building REST API service

- ▶ need to get global tag, luminosity information, etc for many runs
- ▶ built internal REST API service to facilitate the task
- ▶ contributing to similar efforts ongoing in CMS for internal analysis preservation

```
$ curl -s 'http://.../years?year=2015&type=pp' | jq
[
  {
    "year": 2015,
    "type": "pp",
    "lumi_uncertainty": 1.6,
    "luminosity_reference": "https://cds.cern.ch/record/2759951",
    "recid_val": 14210,
    "val_json": [
      {
        "type": "golden",
        "recid": 14210,
        "url": "https://cms-service-dqmdc.web.cern.ch/CAF/certification/Collis..."
      },
      {
        "type": "muon",
        "recid": 14211,
        "url": "https://cms-service-dqmdc.web.cern.ch/CAF/certification/Collis..."
      }
    ],
    "cmssw": "CMSSW_7_6_7",
    "gt_data": "76X_dataRun2_16Dec2015_v0",
    "gt_mc": "76X_mcRun2_asymptotic_RunIIFall15DR76_v1",
    "image_gitlab": "gitlab-registry.cern.ch/cms-cloud/cmssw-docker-opendata/cmssw_7_6_7-slc6_amd64_gcc493",
    "image_dockerhub": "cmsopendata/cmssw_7_6_7-slc6_amd64_gcc493"
  }
]
```

Part III: Data usage examples



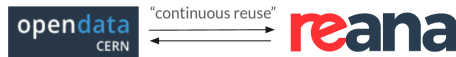
- ▶ containerised data analysis workflows using original research environments
- ▶ data usage examples best document how to work with the published datasets

CMS derived datasets coming with a usage example studying $H \rightarrow \tau\tau$ decays

“Continuous reuse”

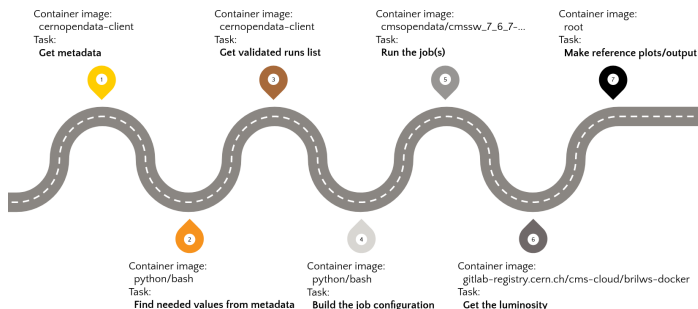
- ▶ periodical execution of data usage examples to detect problems early (with data access, with protocol changes, etc)
- ▶ data production: examples allow to verify the correctness of published provenance information
- ▶ data analysis: examples allow to expose and verify data usage patterns

Analyses								
Name ↑	Last success	Last failure	Last duration	R1	R2	R3	R4	R5
alice-lego-train-test-run	5 hours ago		00:01:35					
alice-pt-analysis	5 hours ago		00:01:10					
atlas-recast	5 hours ago		00:01:10					
cms-dimuon-mass-spectrum	8 days ago		00:01:03					
cms-dimuon-spectrum	5 hours ago		00:01:22					
cms-dimuon-spectrum-nanood	5 hours ago		00:01:52					
cms-h4l	5 hours ago	2 days ago	00:02:02					
cms-h4l-nanood	5 hours ago		00:03:46					
cms-htaufau-nanood	5 hours ago		00:07:08					

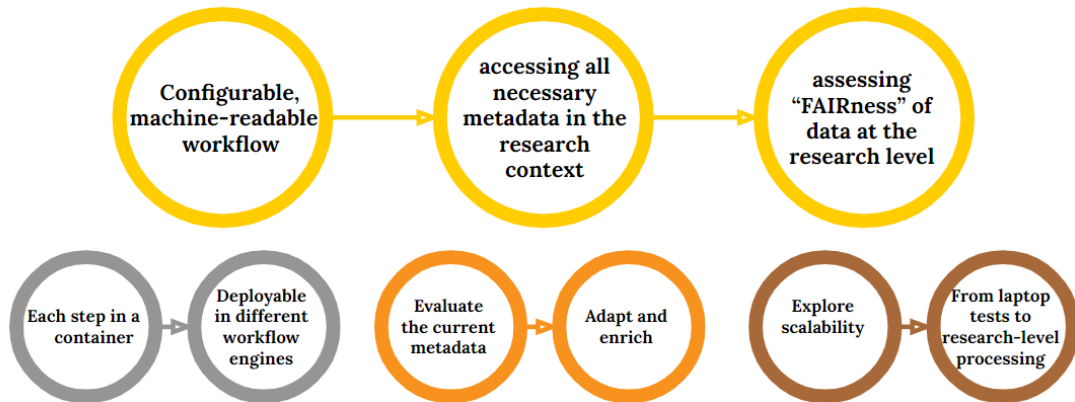


How to ensure research-grade usability of the data?

- ▶ contextual metadata should pass all the knowledge of how to combine the data assets in a meaningful way
- ▶ however, some contextual metadata exist only as part of usage examples and are not directly connected to dataset records
- ▶ need to adapt contextual metadata to make it better retrievable for automated workflows



CMS open data as a testbed



Conclusions

- ▶ CMS has been releasing open data since 2014
 - over eight open data release campaigns
 - over three petabytes of research-grade open data released
- ▶ importance of provenance metadata: how the data came to life
- ▶ importance of contextual metadata: how to connect and use all the data assets correctly
- ▶ importance of data usage examples: actionable knowledge
 - run on original data
 - access condition database
 - expose validated runs, HLT paths, container image names via metadata
- ▶ capturing data knowledge early reduces future data curation detective work



<https://opendata.cern.ch>

