The CMS Open Data workshop at WHEPP XVII: Introduction

January 3rd, 2024

Kati Lassila-Perini
CMS Data preservation and open access coordinator
Helsinki Institute of Physics (Finland)

Welcome!

On behalf of the CMS Open data team

The organizing team of this workshop:

Matt

Julie

Kati

Facilitators:

Aravind

Atul

Ritik

Pruthvi

Mukund

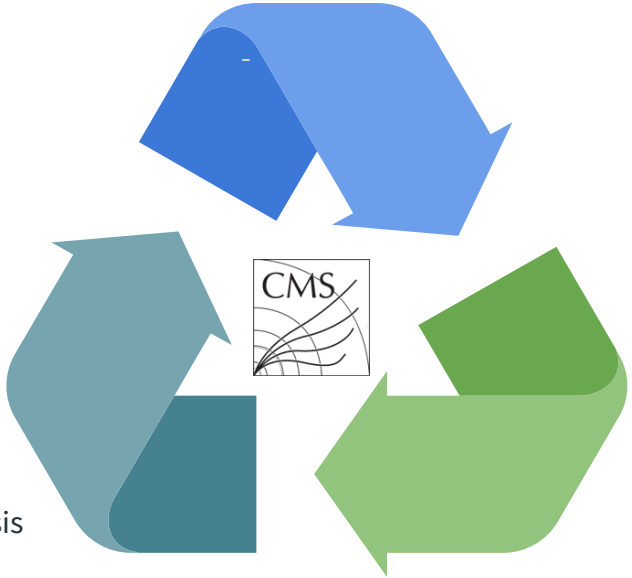
1.

CMS Open data - Why?

Open data as a driving force to data and analysis preservation

Tools:

- software
- environments
- interfaces



Data:

- collision data
- simulations
- additional data for analysis

Knowledge:

- instructions
- actionable examples
- understanding of experimental data

CMS Open data: actual full research-level data - not an “open-data” reduction

But steady publication of LHC data has multiple benefits. First, it encourages prompt archiving, before collective memory fades and knowledge is lost. Second, other scientists can analyse the data while the LHC is still running, testing unconventional strategies and potentially leading to unexpected discoveries, new approaches and fruitful discussions. And third, as a by-product, these scientists can stress test the archiving methods; any deficiencies found are easier to fix now than later. In this way, public collider data can complement the overall LHC research effort. We, therefore, favour a slow but steady approach to full publication of the LHC experiments' data; it is in the best interest of particle physics.

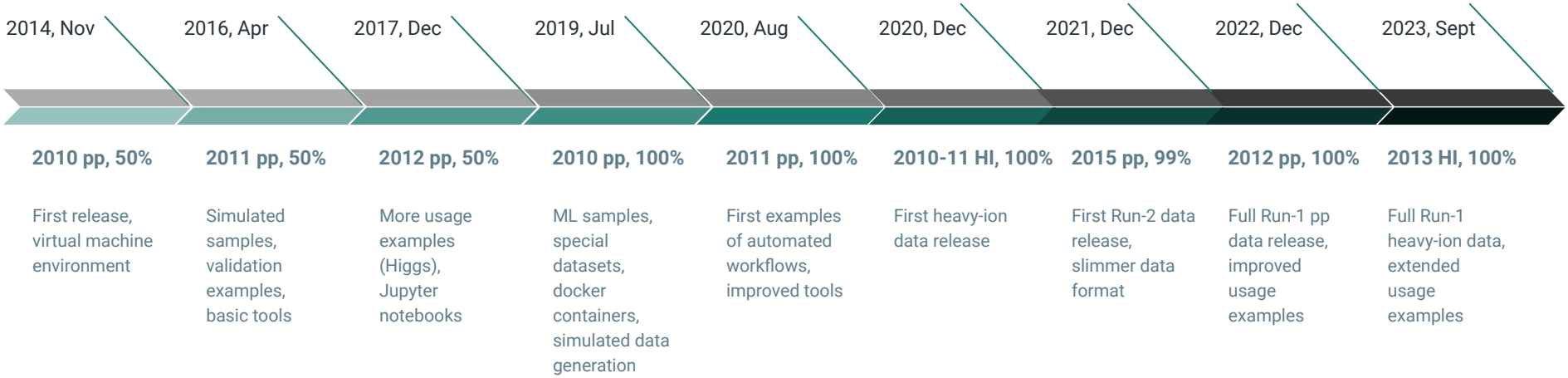
Matthew Strassler, Jesse Thaler
Nature, August 1, 2019
note to the editor

Open data have value only when in use

2.
Release history

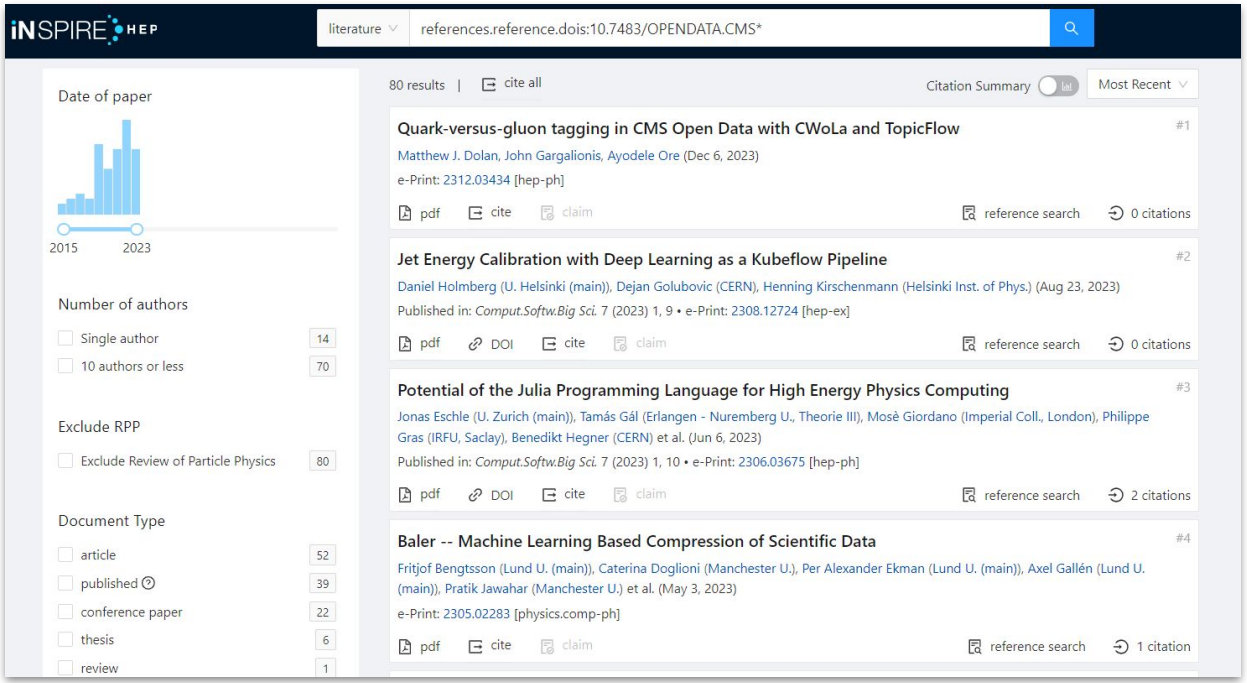
Open data releases since 2014

Release timeline



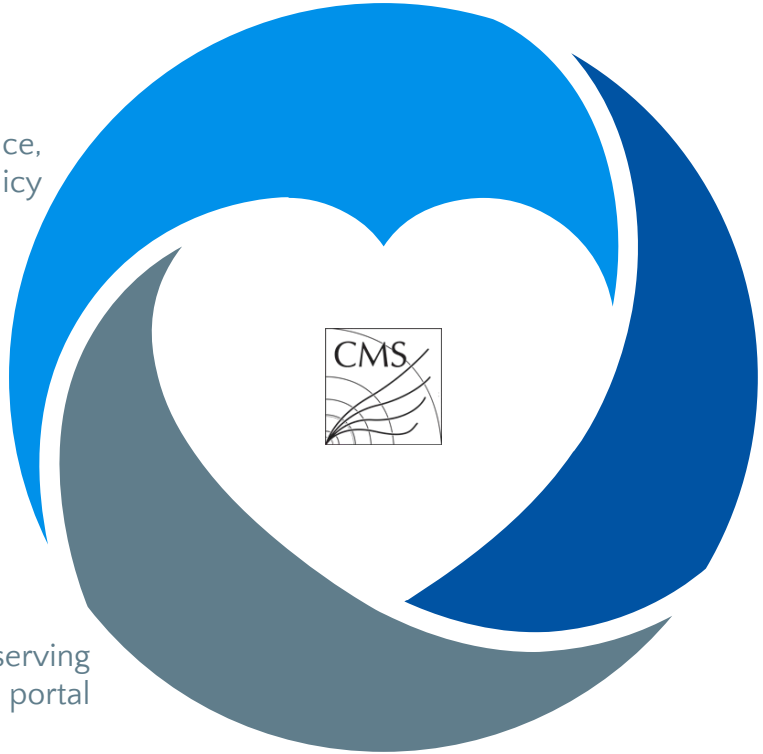
For details, see “CMS Open data” at a workshop Feb/2023 by Julie Hogan

CMS Open data in use



Search (not perfect: does not find all but picks some non-CMS entries)

Positive experience,
model for the CERN policy



Continuous interest,
steady publication rate

Pioneering work for archiving and serving
data through CERN Open data portal

3.

Workshop goals?

What do you expect?
What do we expect?

We made some assumptions

We think that you want to use CMS open data and simulation for physics research.

Therefore, we think you want to understand:

- the basic physics object usage (object access, id, corrections, how to write them out)
- how one can select events and access trigger information
- how to evaluate the luminosity
- the possibilities for large-scale data processing.

In addition, we think you will be interested in

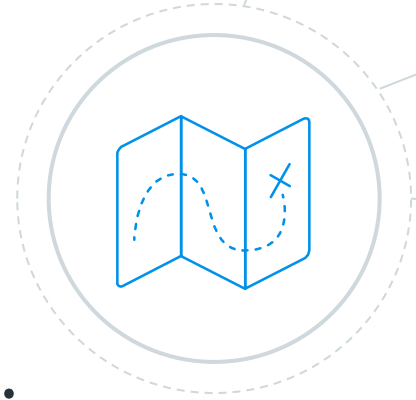
- how to put this all together in an analysis

But that's not all - we get something as well

We want to:

- build a community of users
- remind of https://opendata-forum.cern.ch/
- get understanding of the usage patterns and needs
- get feedback of what is missing in the documentation and tutorial material
- build a proper CMS open data user guide.

Ambitious goals ➟
Do we reach them?



Bear with us:

CMS Open data is always
work in progress

4.

How to get there?

Workshop structure
Working methods

Mandatory “pre-exercises”

Jan 3

17:45-18:10	Welcome to CMS Open Data	Kati Lassila-Perini
18:10-18:15	Orientation to the workshop	
18:15-18:20	Break	
18:20-18:40	Overview of the CMS detector	Helpers
18:40-19:15	Docker container setup and exploration	Helpers

Jan 4

17:45-18:35	ROOT with C++ and Python	Helpers
18:35-18:45	Break	
18:45-19:15	Jupyter & Coffea setup for analysis example	Helpers

Jan 5

17:45-18:15	Check access to TIFR cluster	Helpers
18:15-18:25	Break	
18:25-19:15	Introduction to CMS Physics Objects	Helpers

Pre-exercises

- Importantly, to **set up and test** your working environment before the lessons next week:
  - using CMS open data containers on your own laptop
  - ROOT and other analysis tools
- To give some background information:
  - overview of the CMS detector
  - introduction to physics objects in CMS data
- Scheduled, work on your own pace
  - the facilitators will be present in the room.

Schedule

Jan 6

17:45-17:55	Welcome & Introductions	Julie Hogan
17:55-18:25	Finding CMS Open Data (Lesson)	Julie Hogan
18:25-18:35	Break	
18:35-19:15	Inspecting CMS data files (Activity)	Julie Hogan

Jan 7

17:45-18:25	Event selection (Lesson)	Julie Hogan
18:25-18:35	Break	
18:35-19:15	Event selection (Activity)	Julie Hogan
Bonus Material	Accessing trigger information	
Bonus Material	Advanced tools	

Jan 9

17:45-18:25	Analysis example (Lesson)	Matt Bellis
18:25-18:35	Break	
18:35-19:15	Analysis example (Activity)	Matt Bellis
Bonus Material	Create a "stack plot" histogram	
Bonus Material	Systemics & Statistical interpretation	

Jan 10

17:45-18:25	Analysis scale-up (Lesson)	Julie Hogan
18:25-18:35	Break	
18:35-19:05	Analysis scale-up (Activity)	Julie Hogan
19:05-19:15	Closing	Julie Hogan
Bonus Material	Analysis scale-up additional resources	
Bonus Material	Reinterpreting CMS searches	

Full 4 days of work ahead of us!

Material available from the schedule
A dedicated Mattermost channel in cmsodwswhepp24, see how to subscribe in "Orientation"

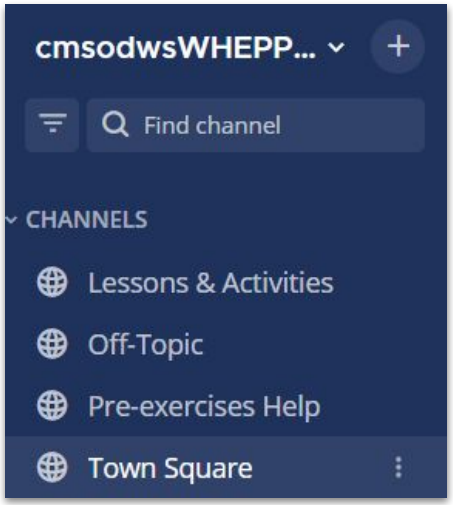
CMS analysis on a HTCondor environment

- You will have the opportunity to learn how to run a CMS open data processing job in real scale on HTCondor queue system using TIFR linux cluster.
- It will be hands-on and you will get a temporary account
  - The hands-on time during the lessons on Wed, Jan 9 is tight
  - Make sure to get everything set up during the intro session on Fri, Jan 5!
- This is new in the CMS Open data workshops - many thanks for helpers for having set it up!!!

- Don’t miss it!

Getting help - live

- In mattermost, choose the channel corresponding to Pre-exercises or Lessons & Activities.
- Do not hesitate to ask!
  - But check if the same question has already been asked.
- Cut and paste the command and the error message
  - If needed, use `some code in line`
  - or ```block of code or output```
  - shift-return for a line break in a message
- Reload the tutorial page every now and then for updates.
- During live lessons
  - In the meeting room, use the mic.



Getting help - live

- The hands-on time during this workshop is short.
  - Make sure to **work through** the pre-exercises.
  - You will make best out of the work if you have problems solved before the hands-on topics next week.
- Do not hesitate to ask: we are there to help you!

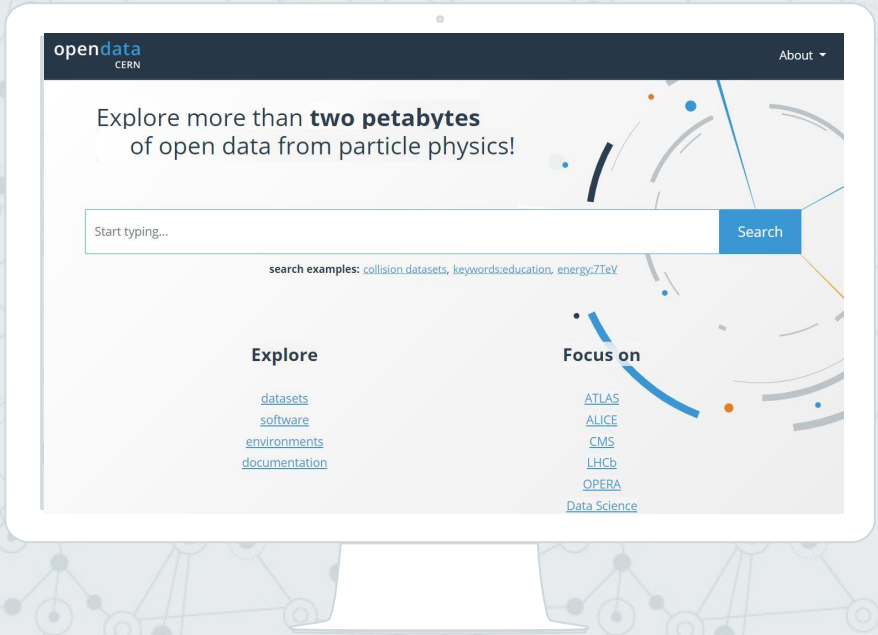
- Please read the instructions carefully
  - WSL2 users: **use the Ubuntu shell**, not Command prompt or Power shell.
  - Mac users: CMSSW container will not work on devices with a M1/M2 chip
  - Suggestions for improvements are most welcome.

Ask! Ask! Ask!

5.

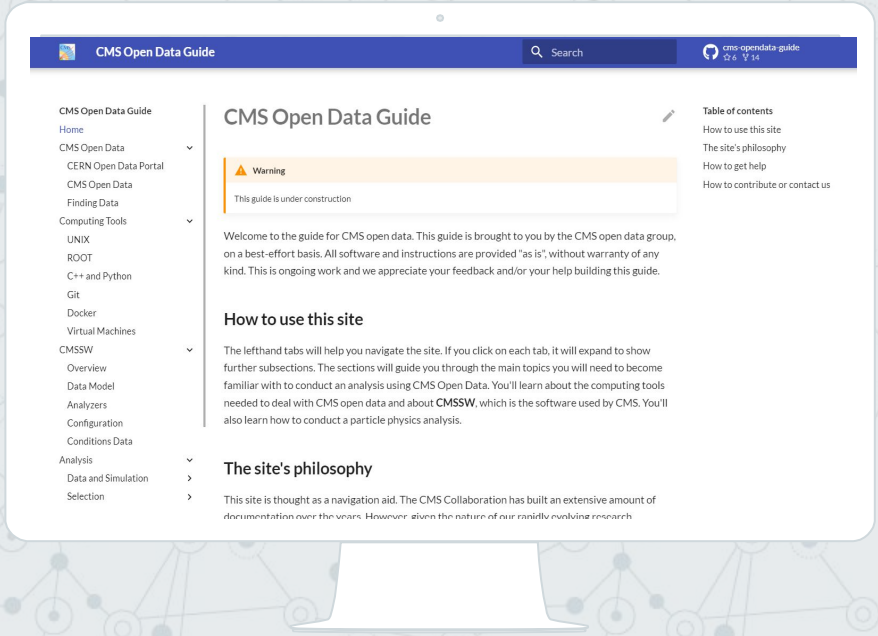
How to get help after?

Information sources
Communication



CERN Open data portal

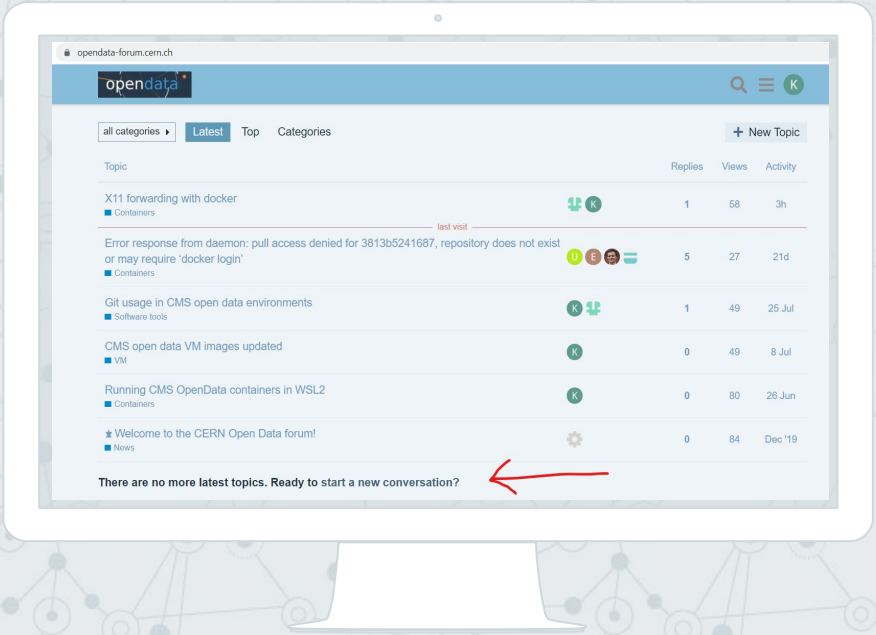
Serves the data, associated analysis artefacts, usage examples



CMS Open data guide

Work in progress, will be completed with the material in this tutorial.

Do you want to help?



CERN Open data forum

Feel free to post questions! Feel free to reply as well!

Most frequently asked questions at this workshop will be added.

Other sources of information

- Open data portal support mail: opendata-support@cern.ch
  - Technical issues
  - Questions to limited audience
- CMS WorkBook and SWGuide
  - Careful: instructions might not correspond to the CMSSW version needed for open data
- CMSSW source code
  - Keep in mind the versioning,
    - for 2011-2012 open data use CMSSW_5_3_X as tag.
    - for 2015 data use CMSSW_7_6_X as tag.

6.

Now, let’s get to work!

Enjoy the workshop!
We’ll love to hear feedback from you
→ Reply to the survey!

Thanks!

Any questions?

Find us in mattermost

Credits

Thanks to the WHEPP organizers for the opportunity!

Thanks to the student facilitators!

Thanks to our colleagues:

- in the DPOA group in CMS
  - all organizers and contributors
- in the CERN Data preservation services
  - CERN Open data portal team, and many other services we rely on

And great thanks to all CMS open data users!

And thanks to SlidesCarnival for this free presentation template