The CMS Open Data workshop: Introduction

July 29, 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

Tom McCauley

Xavier Tintin

Sezen Sekmen

Kati Lassila-Perini

Matt Bellis

Julie Hogan

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.

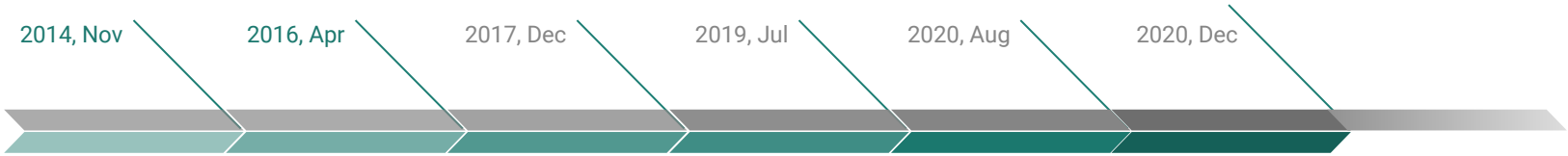
Matthew Strassler, Jesse Thaler
Nature, August 1, 2019
note to the editor

Open data have value only when in use

2.
Release history

Open data releases since 2014

Release timeline



2010 pp, 50%

First release, virtual machine environment

2011 pp, 50%

Simulated samples, validation examples, basic tools

2012 pp, 50%

More usage examples (Higgs), Jupyter notebooks

2010 pp, 100%

ML samples, special datasets, docker containers, simulated data generation tools

2011 pp, 100%

First examples of automated workflows, improved tools

2010-11 HI, 100%

First heavy-ion data release

2015 pp, 99%

First Run-2 data release, slimmer data format

2012 pp, 100%

Full Run-1 pp data release, improved usage examples

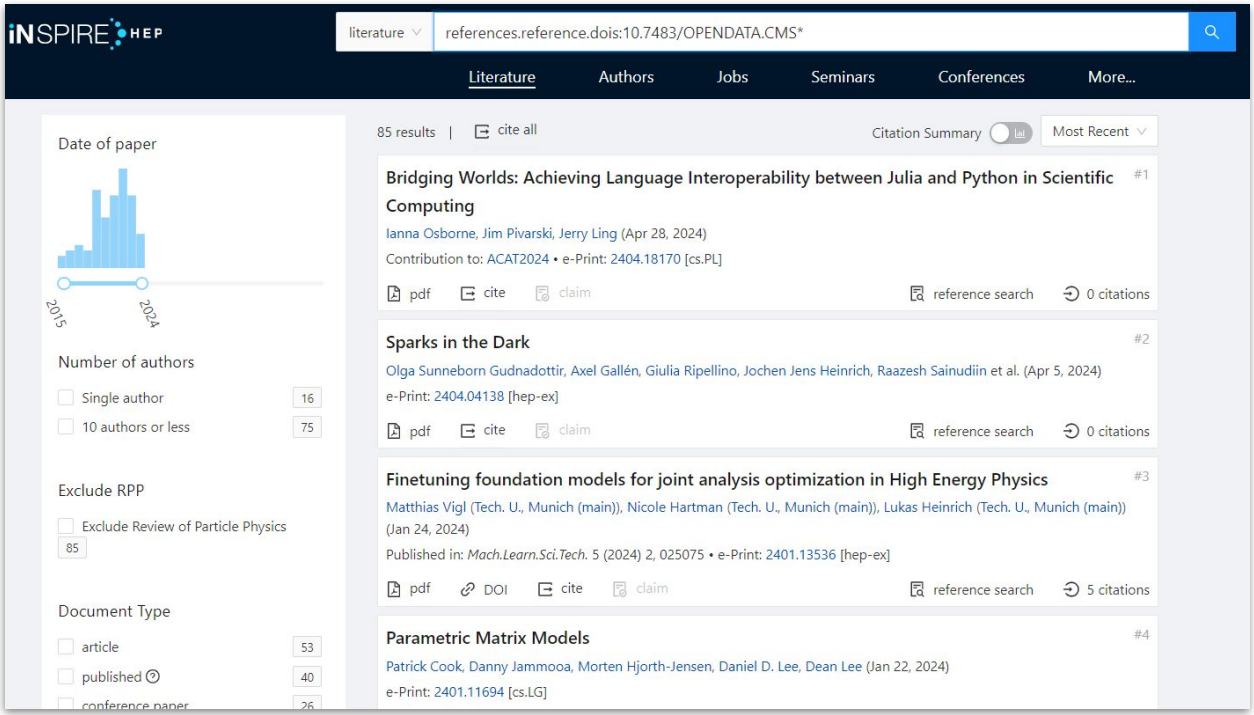
2013 HI, 100%

Full Run-1 heavy-ion data, extended usage examples

2016 pp, 50%

First NanoAOD, updated usage instructions

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

You want to use CMS open data and simulation for physics research.

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
- experimental and statistical uncertainties.

You will be interested in

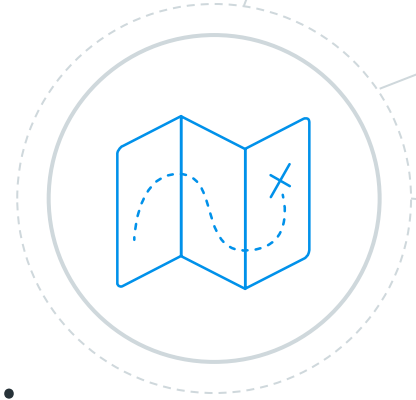
- how to put this all together in an analysis

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

Pre-exercises

Pre-exercises

Completing the required pre-exercises makes full participation in the workshop possible! Submit homework responses as you complete the pre-exercises.

Mandatory	Orientation	
Optional (external lesson)	The Unix Shell	(Windows users: skip install software and follow the "advanced users" instructions to open a shell)
Optional (external lesson)	Version Control with Git	
Optional (external lesson)	Programming with Python	
Mandatory	Docker containers	
Mandatory	Open Data analysis in C++ and Python	
Mandatory	Finding & using open data	

Pre-learning

Pre-Learning

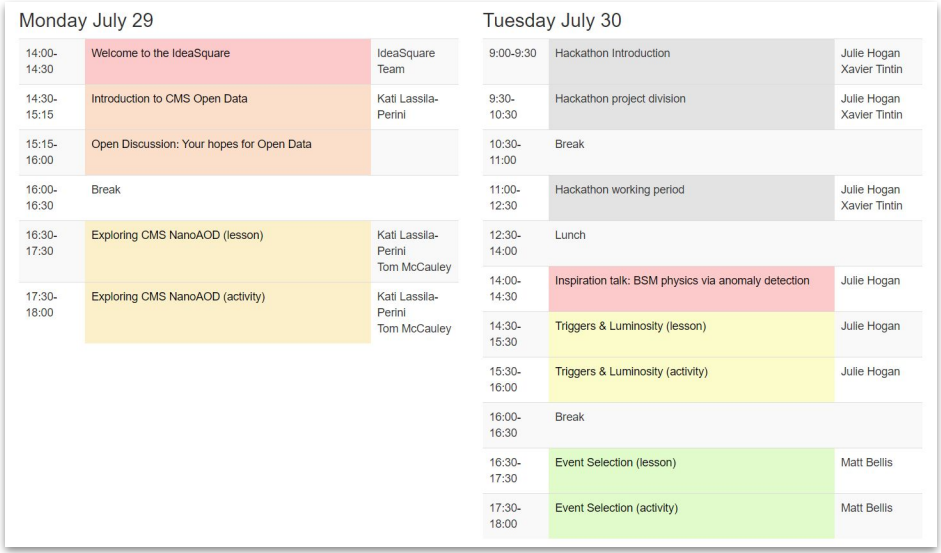
The following concepts are important for using Open Data independently, but will not be taught live during the workshop.

Recommended	Particle Physics Primer	for undergraduates!
Recommended	Overview of the CMS detector	
Recommended	Physics Objects	

Pre-exercises & pre-learning

- Importantly, to **set up and test** your working environment before the lessons
  - using CMS open data containers on your own laptop
  - make sure to have tools to access open data easily
- To give some background information:
  - overview of the CMS detector
  - introduction to physics objects in CMS data
- Not done?
  - let us do it today!

Schedule

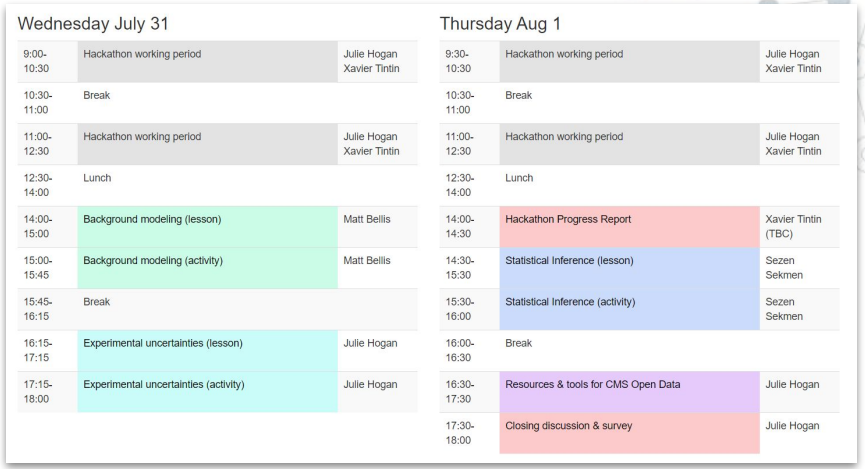


Monday July 29

14:00-14:30	Welcome to the IdeaSquare	IdeaSquare Team
14:30-15:15	Introduction to CMS Open Data	Kati Lassila-Perini
15:15-16:00	Open Discussion: Your hopes for Open Data	
16:00-16:30	Break	
16:30-17:30	Exploring CMS NanoAOD (lesson)	Kati Lassila-Perini Tom McCauley
17:30-18:00	Exploring CMS NanoAOD (activity)	Kati Lassila-Perini Tom McCauley

Tuesday July 30

9:00-9:30	Hackathon Introduction	Julie Hogan Xavier Tintin
9:30-10:30	Hackathon project division	Julie Hogan Xavier Tintin
10:30-11:00	Break	
11:00-12:30	Hackathon working period	Julie Hogan Xavier Tintin
12:30-14:00	Lunch	
14:00-14:30	Inspiration talk: BSM physics via anomaly detection	Julie Hogan
14:30-15:30	Triggers & Luminosity (lesson)	Julie Hogan
15:30-16:00	Triggers & Luminosity (activity)	Julie Hogan
16:00-16:30	Break	
16:30-17:30	Event Selection (lesson)	Matt Bellis
17:30-18:00	Event Selection (activity)	Matt Bellis



Wednesday July 31

9:00-10:30	Hackathon working period	Julie Hogan Xavier Tintin
10:30-11:00	Break	
11:00-12:30	Hackathon working period	Julie Hogan Xavier Tintin
12:30-14:00	Lunch	
14:00-15:00	Background modeling (lesson)	Matt Bellis
15:00-15:45	Background modeling (activity)	Matt Bellis
15:45-16:15	Break	
16:15-17:15	Experimental uncertainties (lesson)	Julie Hogan
17:15-18:00	Experimental uncertainties (activity)	Julie Hogan

Thursday Aug 1

9:30-10:30	Hackathon working period	Julie Hogan Xavier Tintin
10:30-11:00	Break	
11:00-12:30	Hackathon working period	Julie Hogan Xavier Tintin
12:30-14:00	Lunch	
14:00-14:30	Hackathon Progress Report	Xavier Tintin (TBC)
14:30-15:30	Statistical Inference (lesson)	Sezen Sekmen
15:30-16:00	Statistical Inference (activity)	Sezen Sekmen
16:00-16:30	Break	
16:30-17:30	Resources & tools for CMS Open Data	Julie Hogan
17:30-18:00	Closing discussion & survey	Julie Hogan

Full 4 days of work ahead of us!

Material available from the schedule

A dedicated Mattermost channel in cmsodws2024, see how to subscribe in "Orientation"

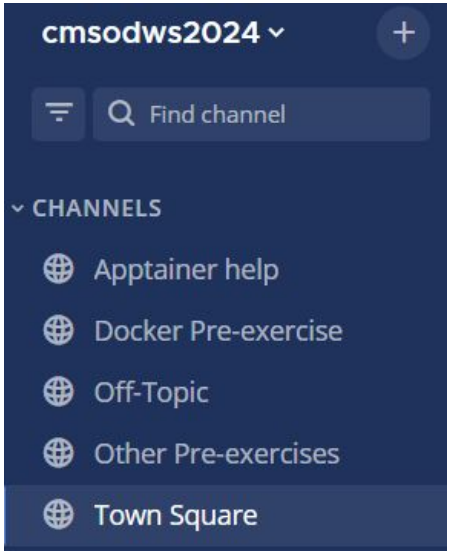
Hackathon

- Your opportunity to get started with some specific topics of your interest.
- Thanks to those of your who replied to the hackathon questionnaire!
- We will offer some suggestions at various levels - more on that tomorrow!

- New in the CMS Open data workshops - many thanks for helpers for having set it up!!!

Getting help - live

- In mattermost, choose the channel corresponding to Pre-exercises or Lessons topic.
- 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:
  - on mornings for hackathon topics
  - scheduled with the lessons on lesson topics (not strictly on the timeslots of the agenda)
- Do not hesitate to ask: we are there to help you!

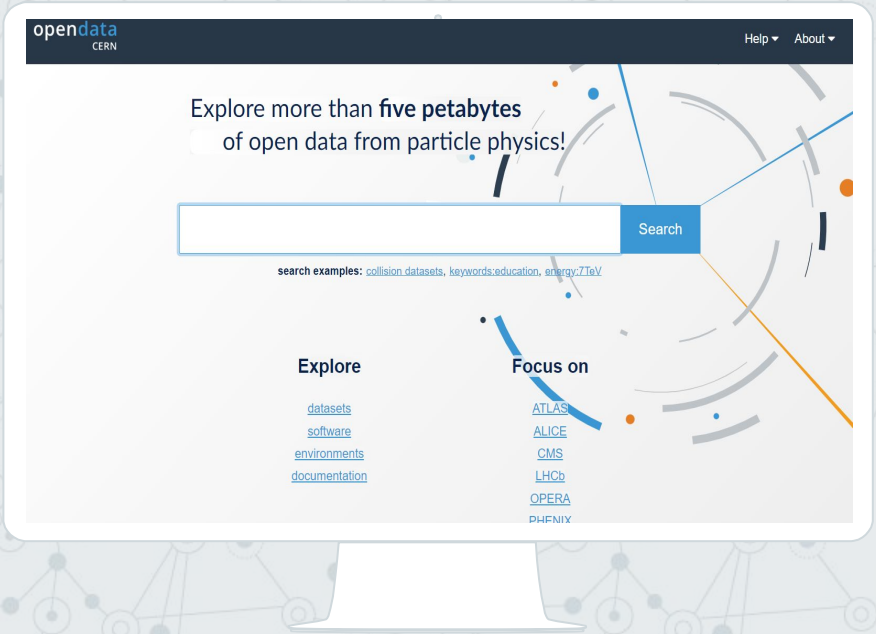
- Make sure to **work through** the pre-exercises.
- Please read the instructions carefully
  - WSL2 users: **use the Ubuntu shell**, not Command prompt or Power shell.
  - Suggestions for improvements are most welcome.

Ask! Ask! Ask!

5.

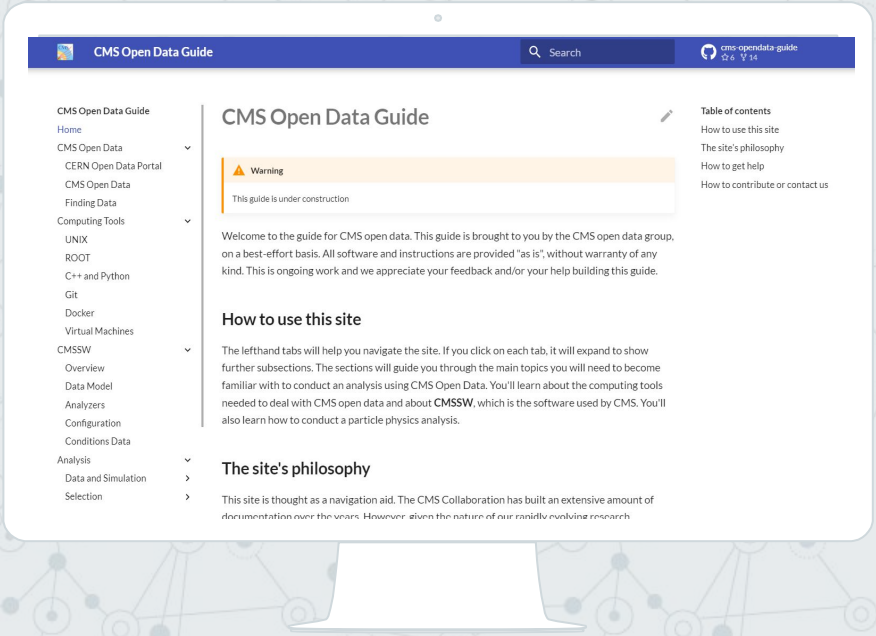
How to get help after?

Information sources
Communication



CERN Open data portal

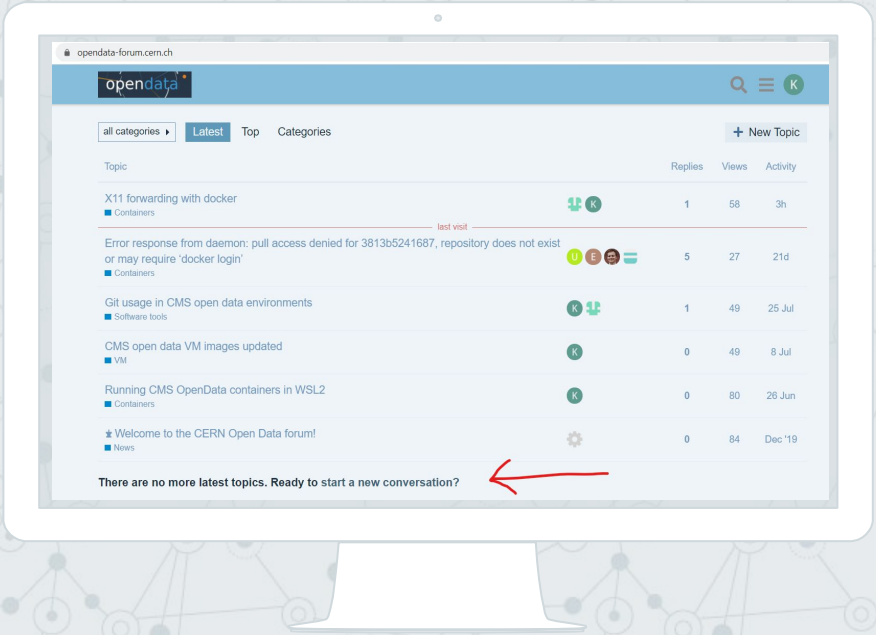
Serves the data, associated analysis artefacts, usage examples



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.
    - for 2016 data use CMSSW_10_6_30 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 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
- Thanks to IdeaSquare for the premises!

And great thanks to all CMS open data users!

And thanks to SlidesCarnival for this free presentation template