

# FAIR4HEP: CMS open data



7 June 2022  
Kati Lassila-Perini  
Helsinki Institute of Physics - Finland  
CMS Data preservation and open access coordinator



# Hello!

I am **Kati Lassila-Perini**

experimental particle physicist

CMS data preservation and open access (DPOA) coordinator

Find me at: [kati.lassila-perini@cern.ch](mailto:kati.lassila-perini@cern.ch)

[@KatiLassila](https://twitter.com/KatiLassila)

1

# CMS Open data - Why?

Open data as a driving force to data and analysis preservation

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

## Before I forget

CMS open data have been a great success

2014, Nov

2016, Apr

2017, Dec

2019, Jul

2020, Aug

2020, Dec

2021, Dec

**2010 pp, 50%**

**2011 pp, 50%**

**2012 pp, 50%**

**2010 pp, 100%**

**2011 pp, 100%**

**2010-11 HI, 100%**

**2015 pp, 99%**

First release, virtual machine environment

Simulated samples, validation examples, basic tools

More usage examples (Higgs), Jupyter notebooks

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

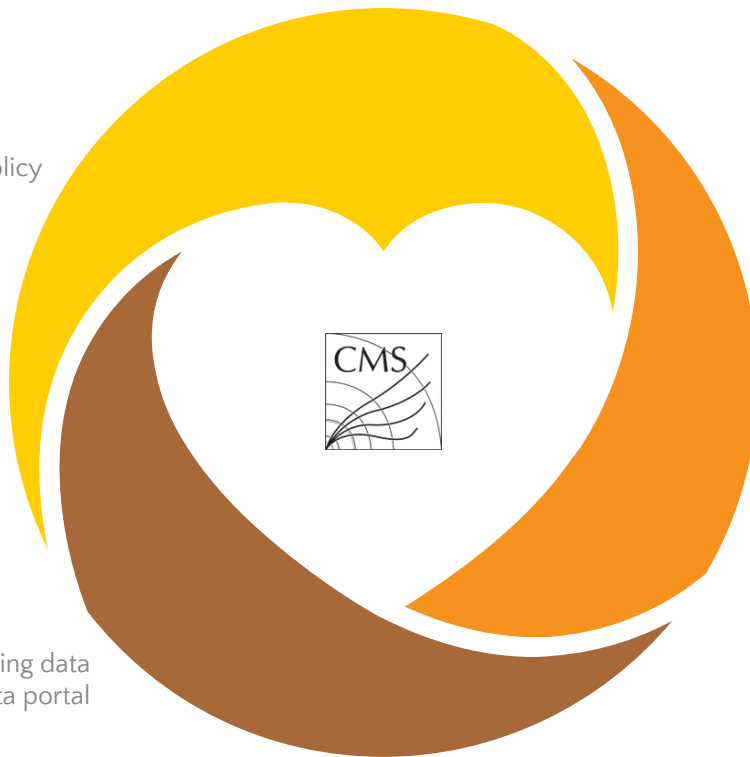
First examples of automated workflows, improved tools

First heavy-ion data release

First Run-2 data release, slimmer data format



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

## CMS Open data - FAIR?

Findable - Accessible - Interoperable - Reusable



## FAIR?

### FINDABLE

From CERN Open Data portal  
(if the search keywords are  
good enough)



F

A

### ACCESSIBLE

XROOTD or direct HTTP  
Command-line tools  
available

Depends.  
Container images provided,  
data formats specific but  
convertible

I

R

Any use is reuse.  
Would be most usefully  
assessed through automated,  
scalable example workflows.

### INTEROPERABLE

### REUSABLE



## **FAIR** is nice, but it is all about usability

- FAIR is often assessed in terms of metadata.
- For complex data, it is not enough!
- Distinguish
  - “direct” metadata – what?
  - “contextual” metadata – how to use, interpret.
    - *“provides a broader understanding by showing how disparate pieces of data relate to each other, placing them into a larger picture.”*

4

# HEP data are complex

Or is it just an excuse?



# What's so complex?

## “Simple” data



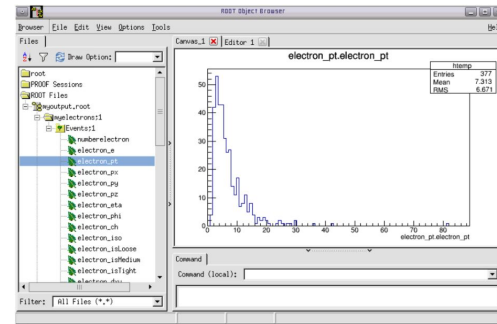
Click  
Select  
View 👍

<https://www.meteoswiss.admin.ch/home/measurement-values.html?param=messwerte-lufttemperatur-10min&station=BLA&chart-hour>

## CMS open data

– properties of particles

- Install docker
- Download image
- Download code
- Compile
- Select data
- Run executable
- Open ROOT
- Select
- View



<http://opendata.cern.ch/docs/cms-getting-started-2015>



# What's so complex?

Why are they making it easy and we are not?

Wind of change:  
slimmer data formats  
python-based data  
science tools...

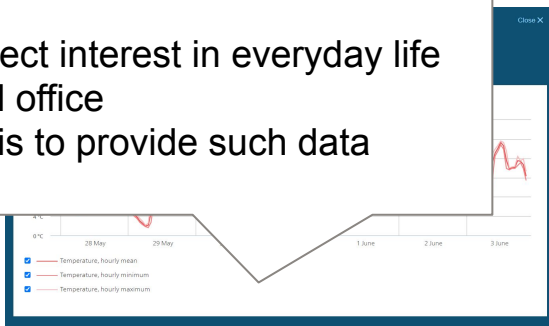
## “Simple” data



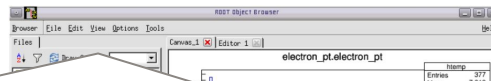
## CMS open data

– properties of particles

- Data of direct interest in everyday life
- Big federal office
- Main task is to provide such data



<https://www.meteoswiss.admin.ch/home/measurement-values.html?param=messwerte-lufttemperatur-10min&station=BLA&chart-hour>



- This simple plot has no direct interest other than illustration
- We are few to work on this, we mostly put the effort of making what we have **usable in research**
- Interactive tools do not easily scale to **research use**
- (We do have a GUI access for few derived samples)

<http://opendata.cern.ch/docs/cms-getting-started-2015>



## What's so complex?

What about the research use?

### “Simple” data



- The fact that we understand a snapshot of these data does not mean that their use in research is easy
- Probably requires multitude of other **different** data from **heterogeneous** sources
- We as particle physicists have no idea of what that takes




<https://www.meteoswiss.admin.ch/home/measurement-values.html?param=messwerte-lufttemperatur-10min&station=BLA&chart-hour>

### CMS open data

- properties of particles

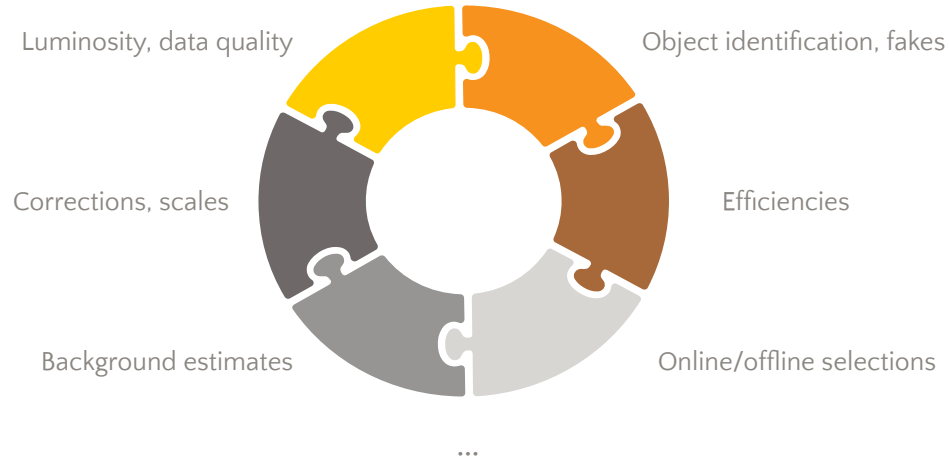


- All necessary information for use of data in research comes **from us**, i.e. single homogenous source
- **We** need to describe and make clear how all that information should be used
- : contextual metadata

<http://opendata.cern.ch/docs/cms-getting-started-2015>



## Contextual metadata - how to get it right





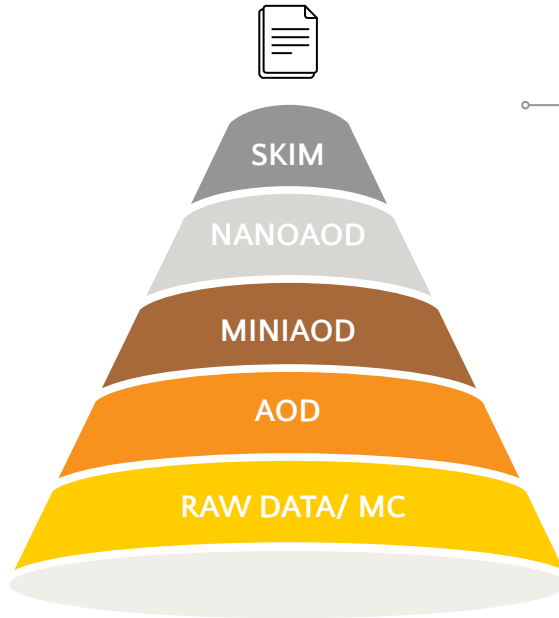


## Contextual metadata - how to get it right

- Teaching/documenting?
  - Open data are CC0: responsibility is on the user.
- We know all this (> 1000 analyses in CMS)
  - Why collecting this for open data is challenging?



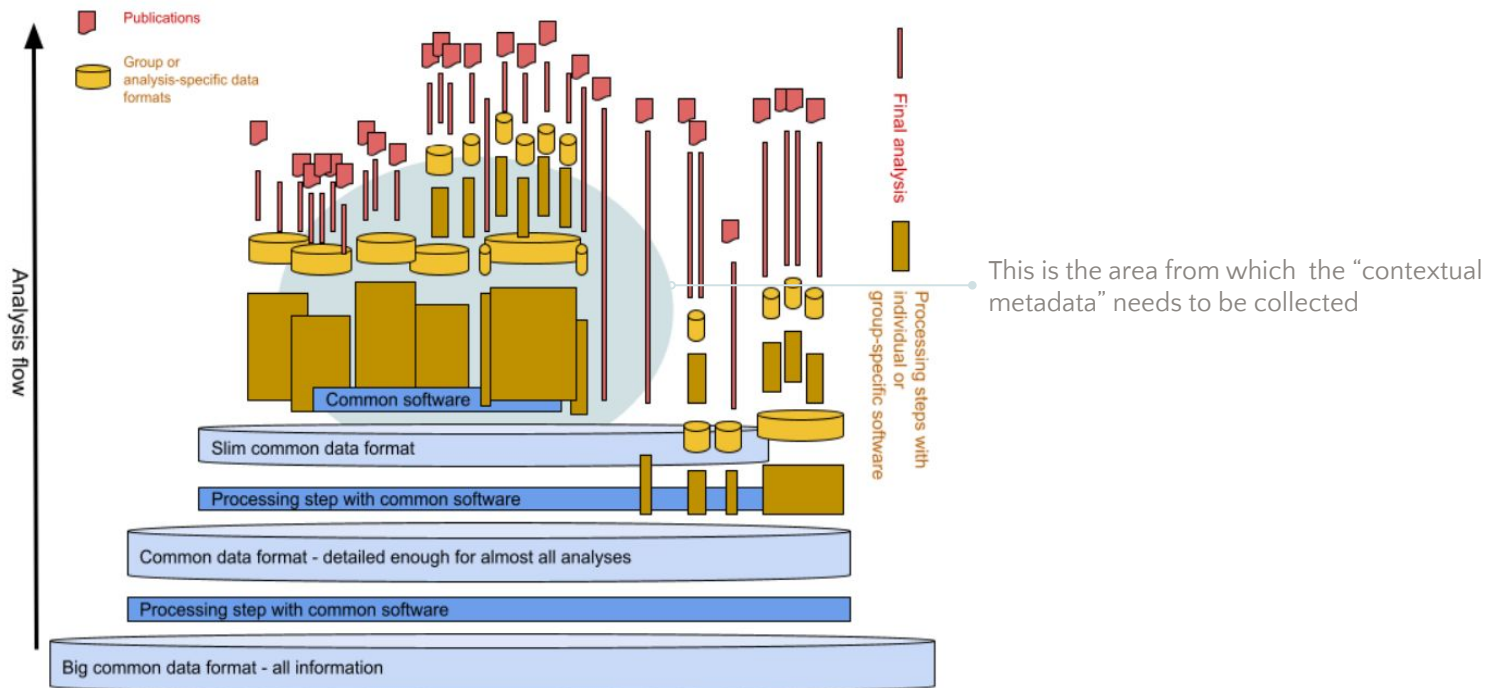
## Data to results - simplified, ideal



- —● Analysis code -> Results
- —● Group/analysis specific skim
- —● Central processing
- —● Central processing
- —● Central processing



# Data to results - in practise





## Why is this so difficult?

---

- Partly because analysis processes are complex.
- But mainly because we, as a community, undervalue:
  - documentation
  - common tools
  - analysis code reuse.

Some further thoughts on this in [a blog](#).



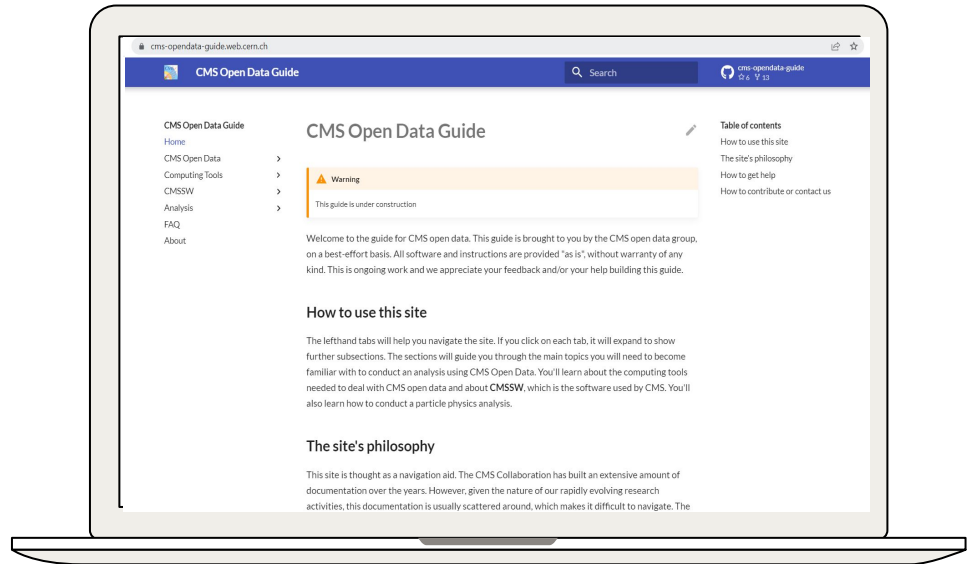
# But we are getting there

## CMS Open Data workshops

Bridges the technical gap between the scientific creativity of an external analyst and the nuts-and-bolts details of a full analysis with CMS open data.

## CMS Open Data user guide

Expands the short, topical guide pages on the Open data portal and aims to be a navigation aid to scattered documentation



5

## Small things matter

Usability must be considered through how it is experienced by the external users and not through how we think it



## It's not what you say, it is what others hear

### Skills

People from different backgrounds and with different ages have different skills.

Do not assume, and make it safe to ask. Overdocumenting is not a shame.

### Tools

We are not in the mainstream with ROOT and C++. Users are familiar with other tools.

Test the usability, from copy-paste of commands to download times.

### Knowledge

Pass knowledge in a usable form, with explicit, working code examples.

Best with workflows understandable to humans and readable by machines.



## Example: OS of workshop participants

	<b>Linux</b>	<b>MacOS</b>	<b>Windows (WSL2)</b>
2021	41%	34%	25%
2022 (as of today)	46%	28%	26%



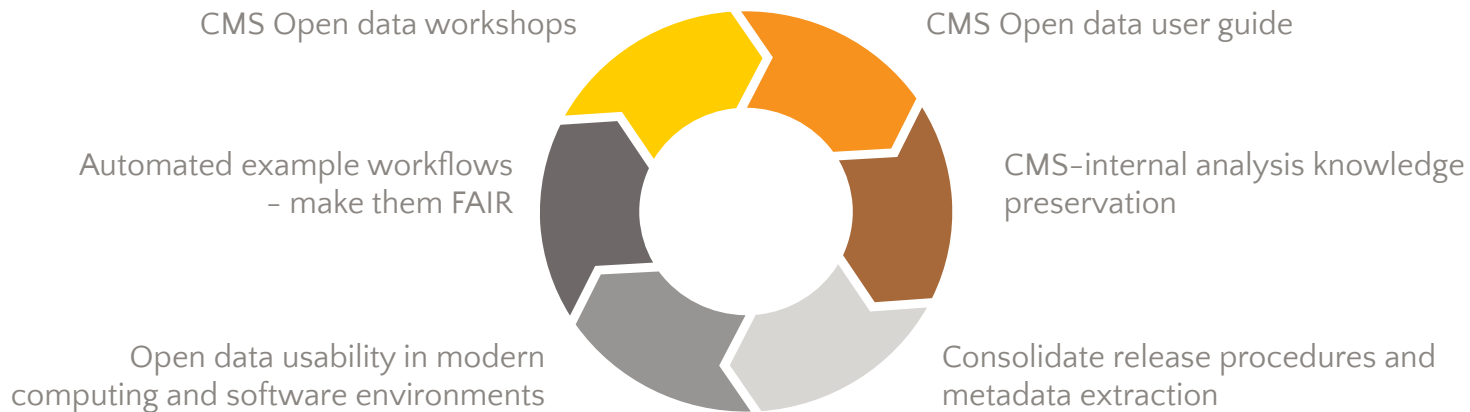
6

## Outlook and plans

When preserving data we certainly need to look back, but most importantly, keep looking forward



## What's on in CMS DPOA?





---

# Thanks!

*Any* **questions** ?



## Credits

---

- Thanks to my colleagues
  - in CMS and, in particular, in the DPOA group
    - Clemens Lange, Edgar Carrera, Lara Lloret, Achim Geiser and many others
  - in the CERN Data preservation services
    - Open data portal and ReANA teams, CAP team, and many other services that we rely on
- Great thanks also to all CMS open data users!