LHC perspective with CMS open data

Belle II Data preservation workshop - 7 Sept 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

@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.

“

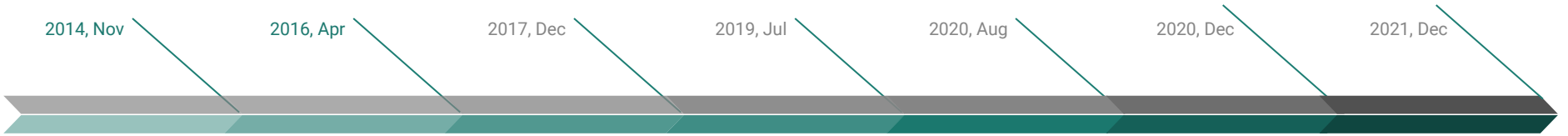
Matthew Strassler, Jesse Thaler
Nature, August 1, 2019
note to the editor

Open data have value only when in use

2

Before I forget

CMS open data have been a great success



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

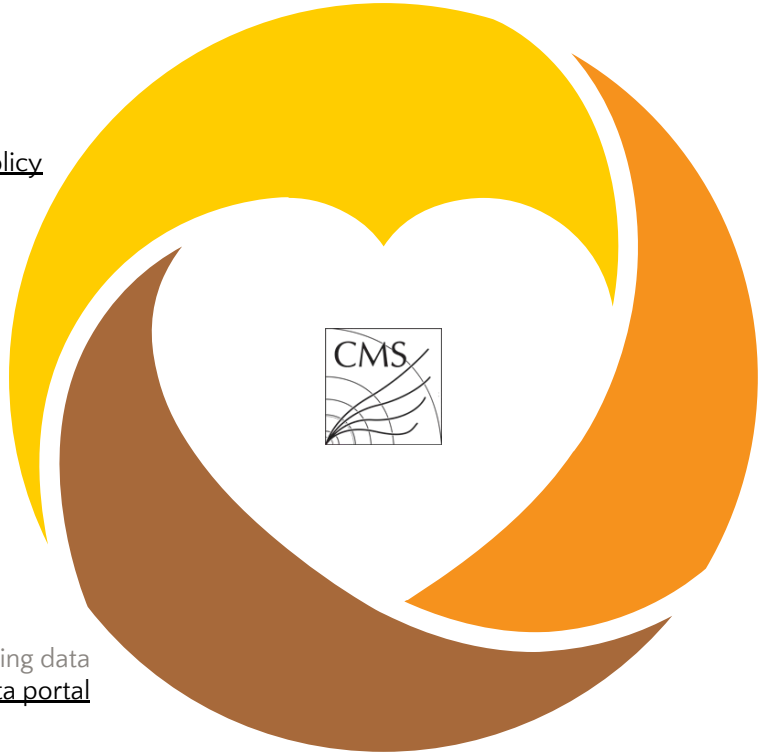
2010-11 HI, 100%

First heavy-ion data release

2015 pp, 99%

First Run-2 data release, slimmer data format

Positive experience, model for the CERN policy



Continuous interest, steady publication rate

Pioneering work for archiving and serving data through the 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)

ACCESSIBLE

XROOTD or direct HTTP

Command-line tools available

Depends.
Container images provided, data formats specific but convertible

INTEROPERABLE

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

REUSABLE

FAIR is nice, but it is all about usability

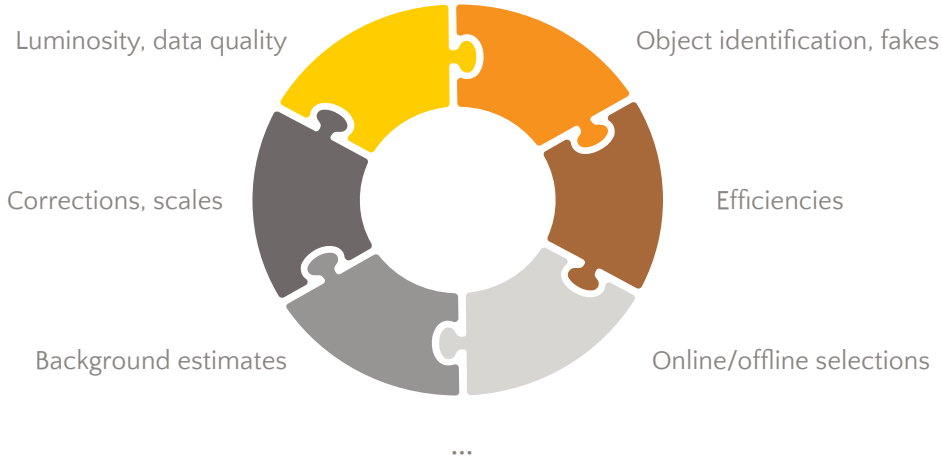
- FAIR is often assessed in terms of metadata.
- For complex data, it is not enough!
- Distinguish
  - “direct” metadata – what?
  - “contextual” metadata – how to use, interpret.
    - *“provides a broader understanding by showing how disparate pieces of data relate to each other, placing them into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

4

HEP contextual metadata

Collecting “direct” metadata from internal sources requires work, but the main challenge is the contextual metadata

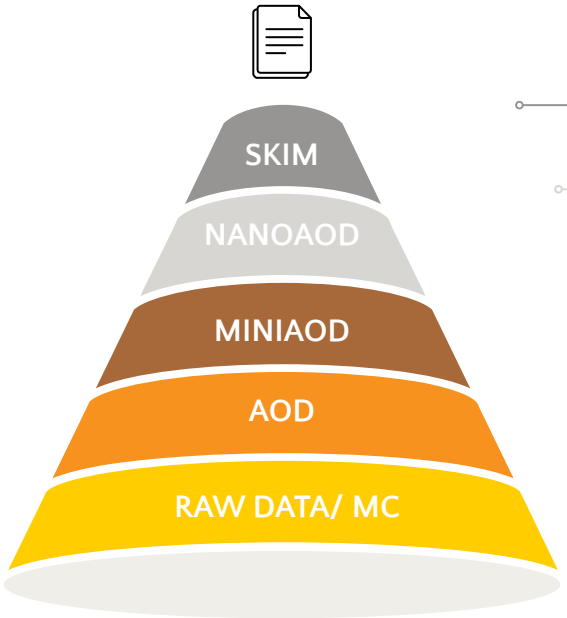
Contextual metadata - how to get it right



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

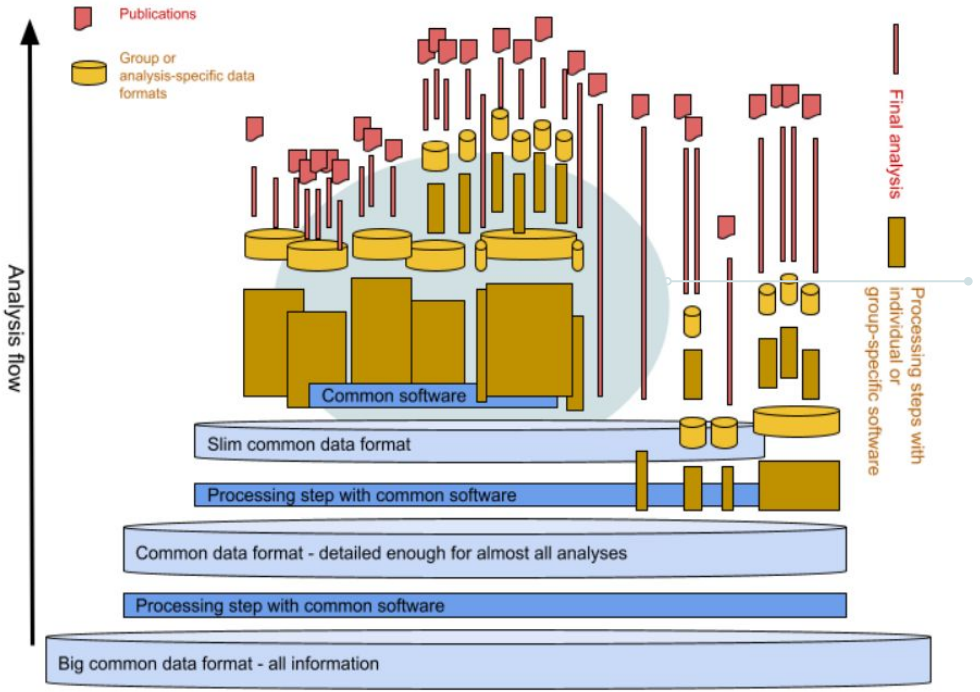
Group/analysis specific skim

Central processing

Central processing

Central processing

Data to results - in practise



This is the area from which the “contextual metadata” needs to be collected

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.

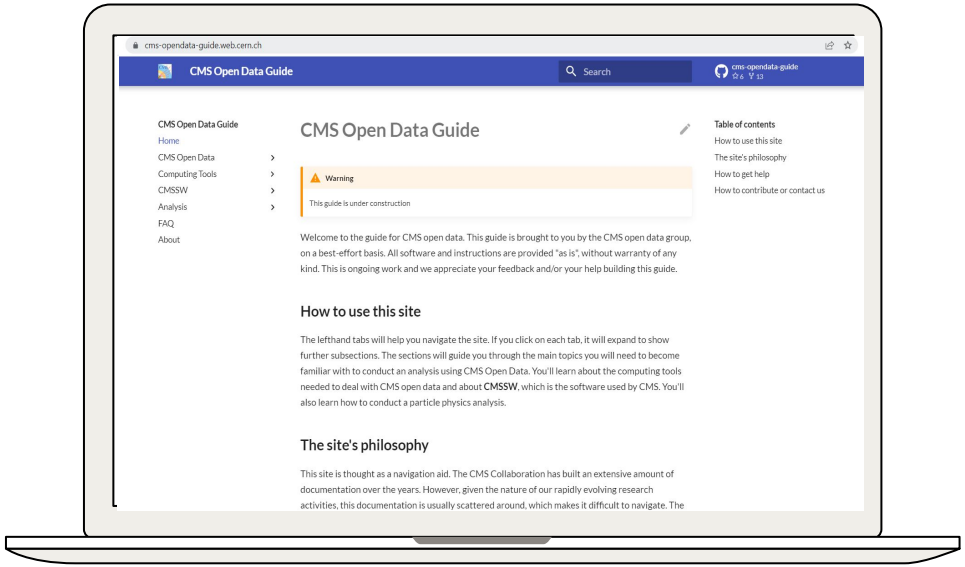
But we are getting there

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

Consider usability through how it is experienced by the external users and not through how you see 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.

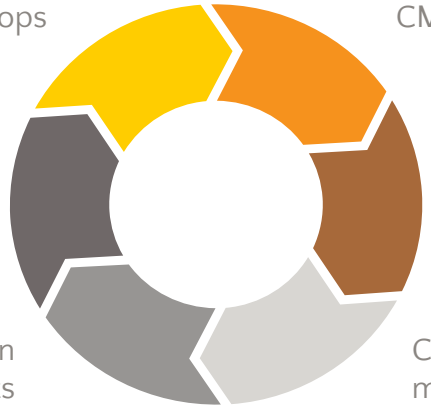
Best with workflows understandable to humans and readable by machines.

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?



7

Food for thought

From my personal experience with CMS open data

Some observations - if you allow me

Your **stress testing** happens very late (timeline in Rec. 4):

- Real test of usability only by external users
  - without experiment-specific resources other than those provided in the open data distribution
- No expertise left to fix and complete open data distribution.

How will you **distribute**, cfr. CERN Open Data portal?

Value and goals of open data:

- beyond "outreach" (i.e. with you as an active partner)
- aim for enabling doing science on them (**without you**)!

Kudos for your efforts for **reproducible analysis workflows** (Rec. 1&2)!!

- If successful (i.e. knowledge preserved), they help you overcome the challenges with the late stress testing.

Thank you!

Any questions ?

Credits

- Thanks to my colleagues
  - in CMS and, in particular, in the DPOA group
    - Clemens Lange, Edgar Carrera, 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!

And thanks to SlidesCarnival for this free presentation template