Open data - how to?
Learning from CMS open data

DK ALM Open data workshop - 16-17 Jan 2023

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

https://sciencemastodon.com/@KatiLassila

1 CMS open data

1.1

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.

From CMS open data users:

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

Open data have value only when in use

1.2

CMS Open data – FAIR?

Findable – Accessible – Interoperable – Reusable

FAIR? My interpretation…

FINDABLE

Do you know where to look for them?

Can you find what you need?

ACCESSIBLE

Can you download them?

Are they in some common format?
Do you have the tools to open the data files?

INTEROPERABLE

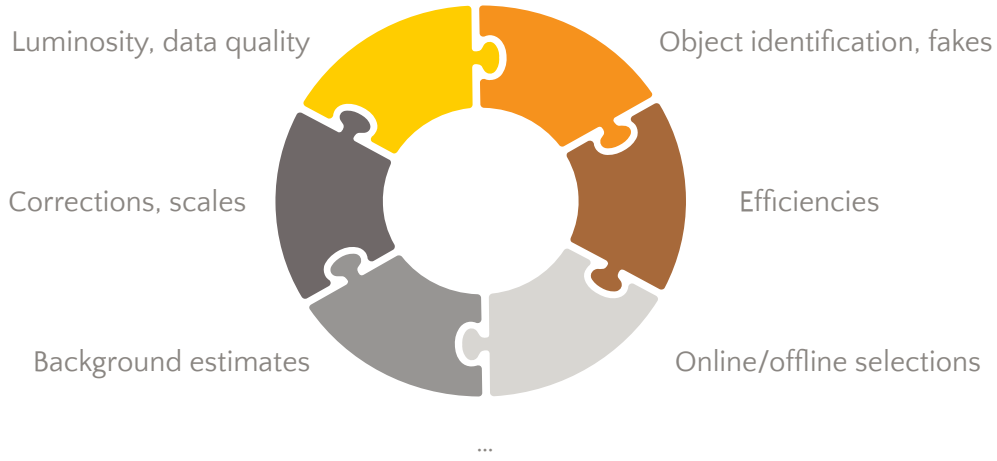
Do you know how to use?
Can you make new research with them?

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
  - “content” metadata – what?
  - “provenance” metadata – from where?
  - “contextual” metadata – how to use, interpret?

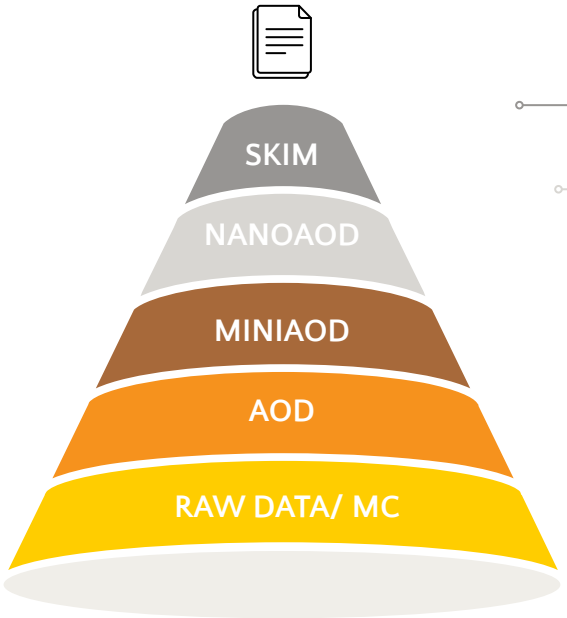
Contextual metadata - how to get it right



Contextual metadata - how to get it right

- We know all this (> 1000 analyses in CMS)
  - But conveying this to open data users is challenging!
- Teaching/documenting?
  - Open data are CC0: responsibility remains on the user.

Data to results - simplified, ideal



Analysis code -> Results

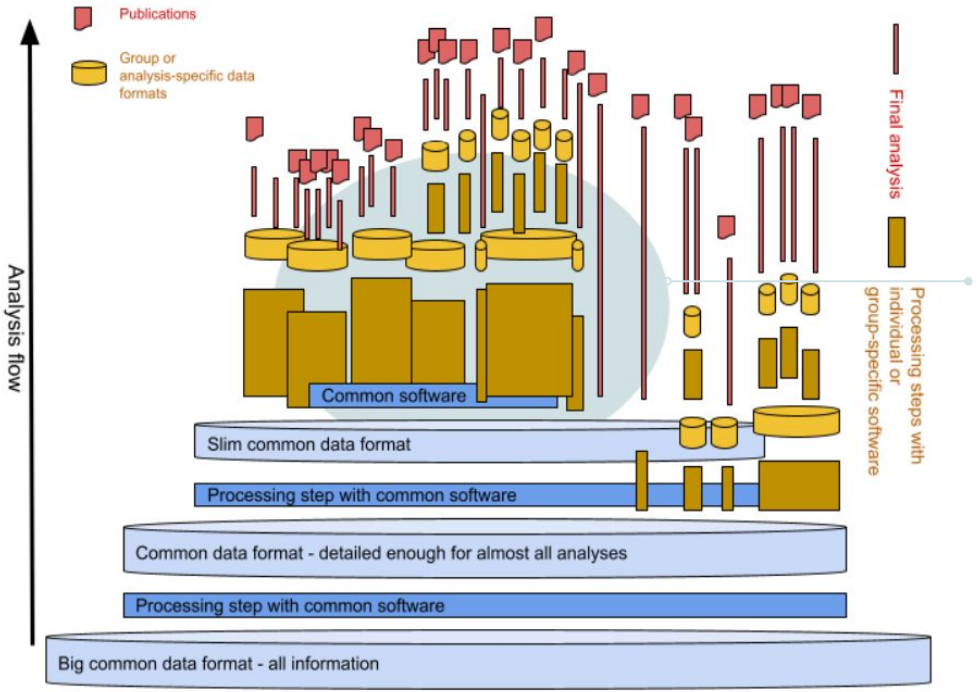
Group/analysis specific skim

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?

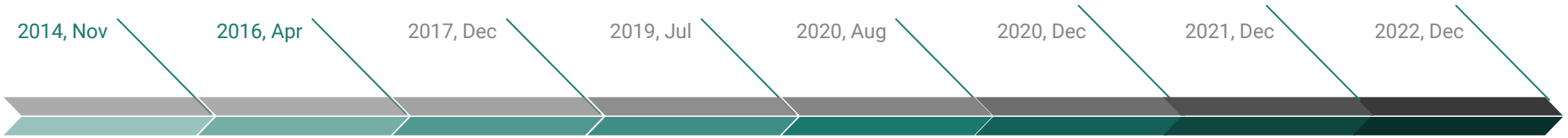
- Partly because analysis processes are complex.
- But mainly because we, the academic community, undervalue:
  - documentation
  - common tools
  - analysis code reuse.

Some further thoughts on this in a blog.

1.3

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

2012 pp, 100%

Full Run-1 pp data release, improved usage examples

Positive experience,
model for the CERN open data policy

Continuous interest,
steady publication rate

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

2

You and open data

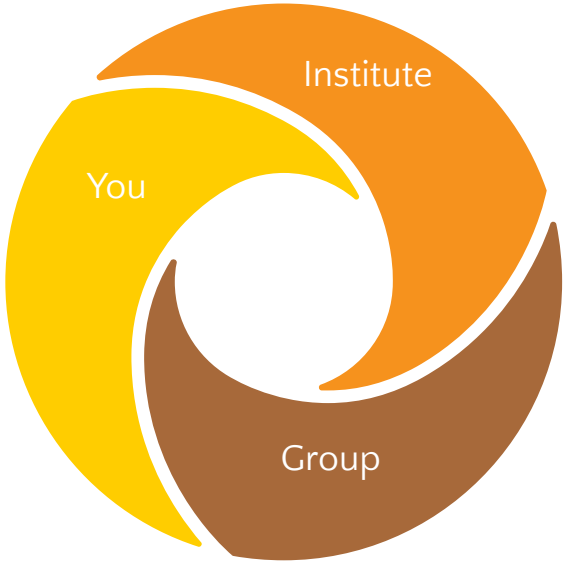
Three points before starting the hands-on part

2.1

Roles & agreements

Individual researcher, collaboration/group, institutes/institutions

Open science - roles



Infrastructure
Resources
Guidelines

Knowledge preservation
Data and software skills

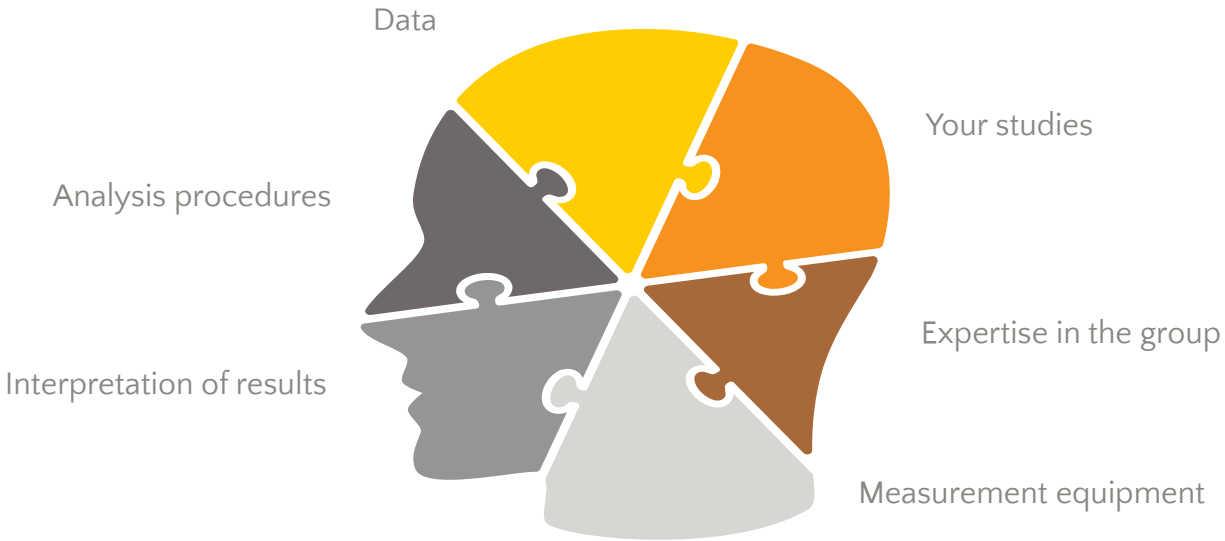
Policies
Agreements
Practices

2.2

Open ≠ Simple

Making data public does not make them simple

Research is a complex process



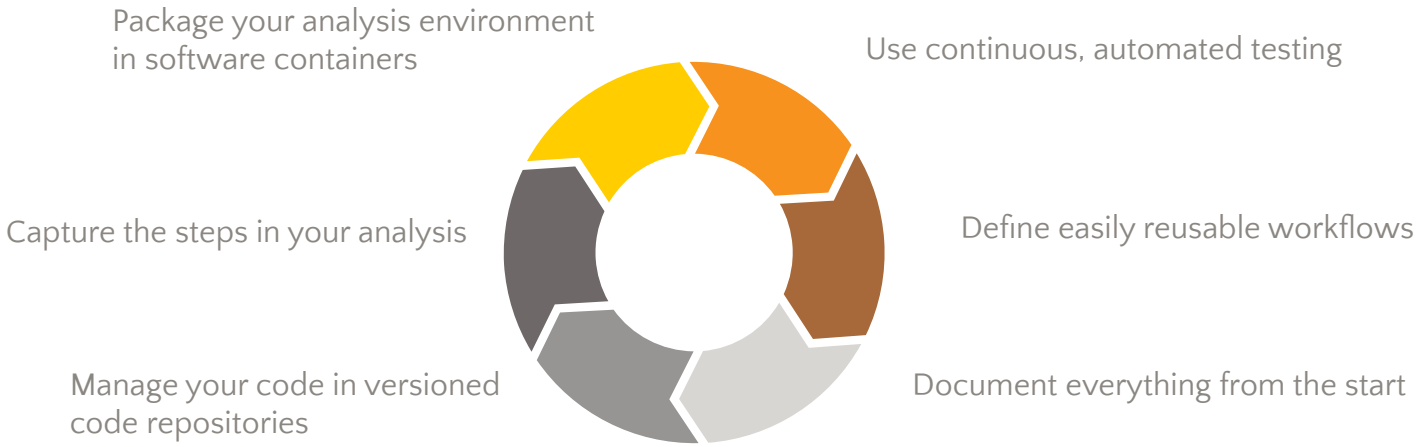
You will not be able to convey all this with open data

2.3

Open ⇎ Reusable

Making data open is necessary but not sufficient for their reusability

Efforts are needed for usability



Best practices that you need to learn for open data will soon pay off: for yourself, for your group, and eventually, for open science!

Thank you!

Any questions ?