CMS Open data - process, challenges and plans

HSF Data Analysis Working Group - 3 July 2023

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

1

CMS Open data

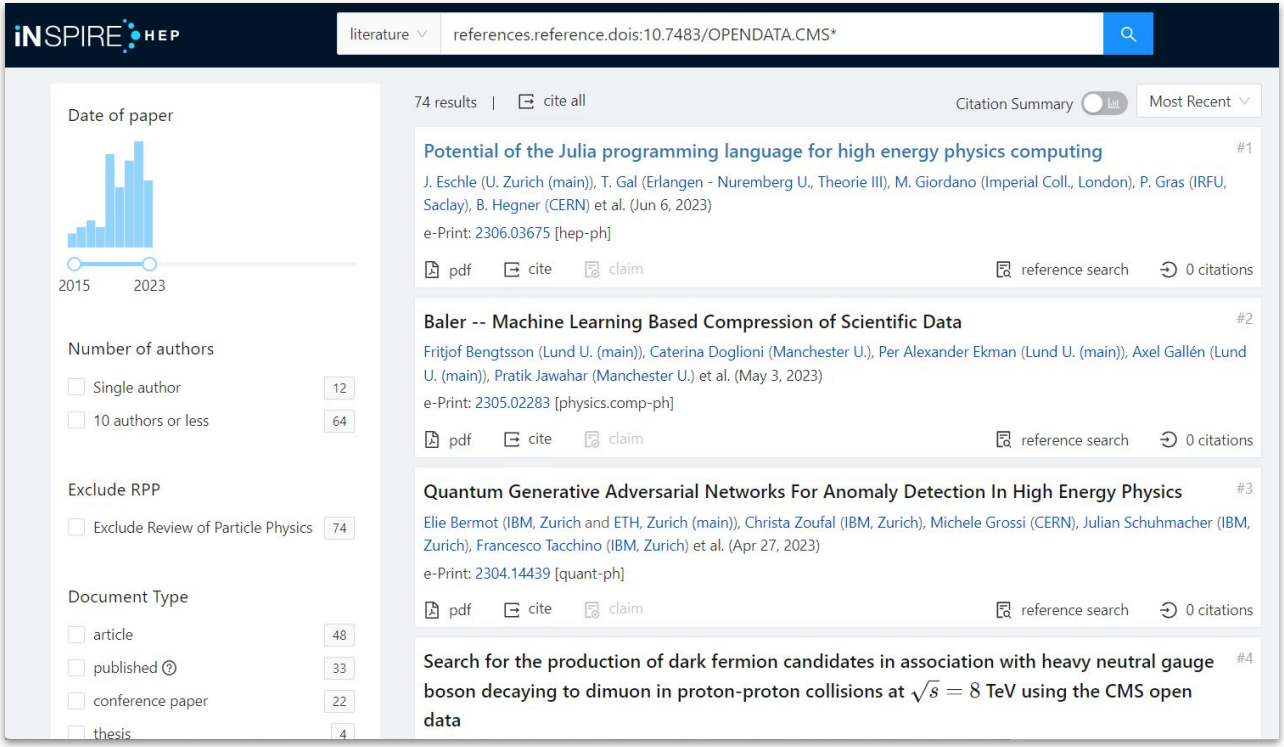
Continuous releases since 2014

- open data appreciated and in use

Release timeline

2014, Nov	2016, Apr	2017, Dec	2019, Jul	2020, Aug	2020, Dec	2021, Dec	2022, Dec
2010 pp, 50%	**2011 pp, 50%**	**2012 pp, 50%**	**2010 pp, 100%**	**2011 pp, 100%**	**2010-11 HI, 100%**	**2015 pp, 99%**	**2012 pp, 100%**
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	Full Run-1 pp data release, improved usage examples

For details, see "CMS Open data" at a recent workshop by Julie Hogan, DPOA co-convener



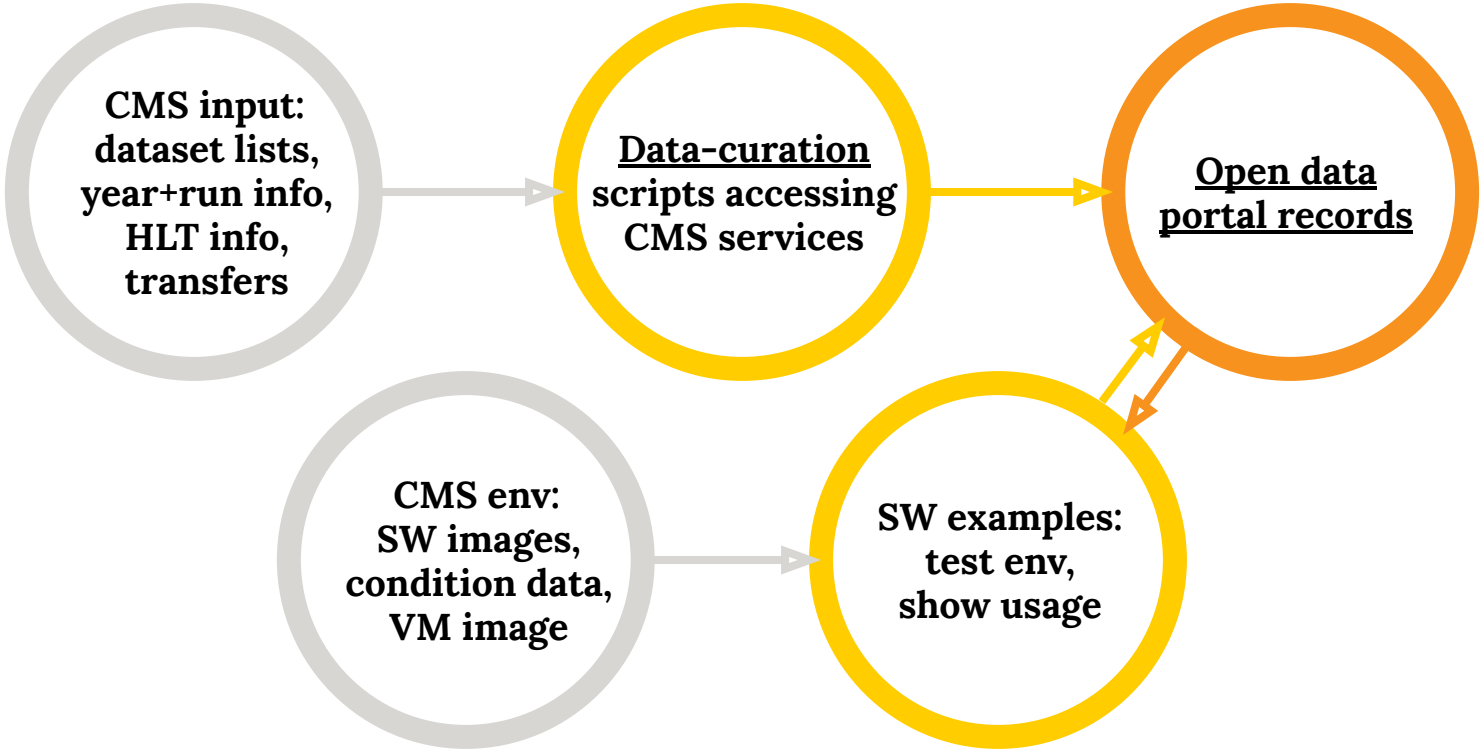
CMS open data are actively in use – no problems have ever arisen!
Search inspire

2

The process

What CMS needs → what external users need

Release procedure



Contextual metadata: HOW TO USE?

Dataset characteristics

76523854 events. **1607** files. **1.5 TB** in total.

System details

Recommended global tag for analysis: 76X_dataRun2_16Dec2015_v0

Recommended release for analysis: CMSSW_7_6_7

How were these data selected?

Events stored in this primary dataset were selected because of the presence of a high scalar sum of the jet transverse momenta (HT); or at least one or two energetic jets.

Data taking / HLT

The collision data were assigned to different RAW datasets using the following HLT configuration.

Data processing / RECO

This primary MINIAOD dataset was processed from the RAW dataset by the following step:

Step: RECO

Release: CMSSW_7_6_3

Global tag: 76X_dataRun2_v15

Configuration file for RECO step reco_2015D_JetHT

HLT trigger paths

The possible HLT trigger paths in this dataset are:

HLT_AK8DiPFJet250_200_TrimMass30_BTagCSV0p45

HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV0p45

HLT_AK8PFHT600_TrimR0p1PT0p03Mass50_BTagCSV0p45

Content metadata: WHAT?

Provenance metadata: FROM WHERE?

Context: environment software

Context: trigger path

Context: usage instructions

Context: validated data selection

Dataset characteristics

76523854 events. **1607** files. **1.5 TB** in total.

System details

Recommended global tag for analysis: 76X_dataRun2_16Dec2015_v0

Recommended release for analysis: CMSSW_7_6_7

How were these data selected?

Events stored in this primary dataset were selected because of the prese
or two energetic jets.

Data taking / HLT
The collision data were assigned to different RAW datasets using the foll

Data processing / RECO
This primary MINIAOD dataset was processed from the RAW dataset by
Step: RECO
Release: CMSSW_7_6_3
Global tag: 76X_dataRun2_v15
Configuration file for RECO step reco_2015D_JetHT

HLT trigger paths
The possible HLT trigger paths in this
HLT_AK8DiPFJet250_200_TrimMass3
HLT_AK8DiPFJet280_200_TrimMass3
HLT_AK8PFHT600_TrimR0p1PT0p03

JetHT primary dataset in MINIAOD format from RunD of 2015 (/ 16Dec2015-v1/MINIAOD)

/JetHT/Run2015D-16Dec2015-v1/MINIAOD, CMS collaboration

Cite as: CMS collaboration (2021). JetHT primary dataset in MINIAOD format fro nD of 201 HT/Run2015
Open Data Portal. DOI:10.7483/OPENDATA.CMS.IDN0.S11Z

Dataset | Collision | CMS | 13TeV | CERN-LHC

Description

JetHT primary dataset in MINIAOD format from R nD of 2015. Run period from run nu 030 to 260627.

The list of validated runs, which must be pplied to all analyses, either with the full validation or for an analysis requ

Validated runs, full validation

How can you use these data?

You can access these data through the CMS Open Data container or the CMS Virtual Machine. See the instructions for setting up one of the two alternative environments and getting started in

Running CMS analysis code using Docker

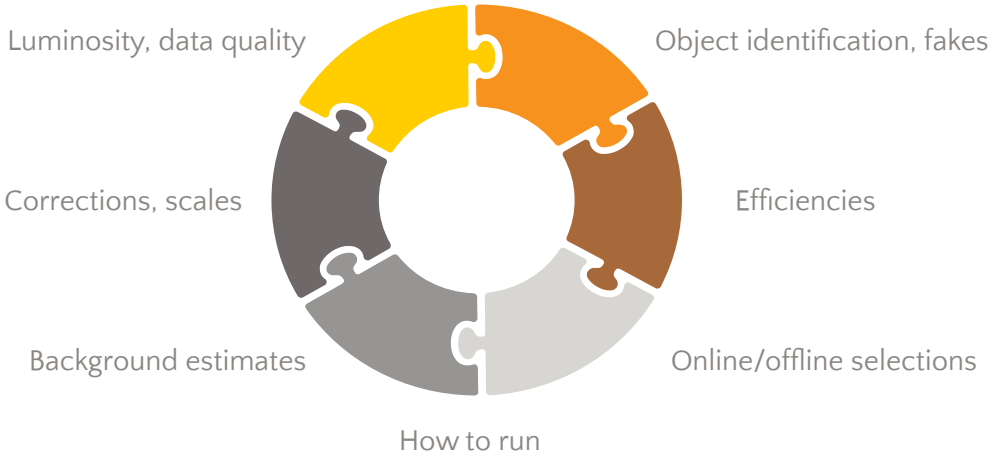
How to install the CMS Virtual Machine

Getting started with CMS open data

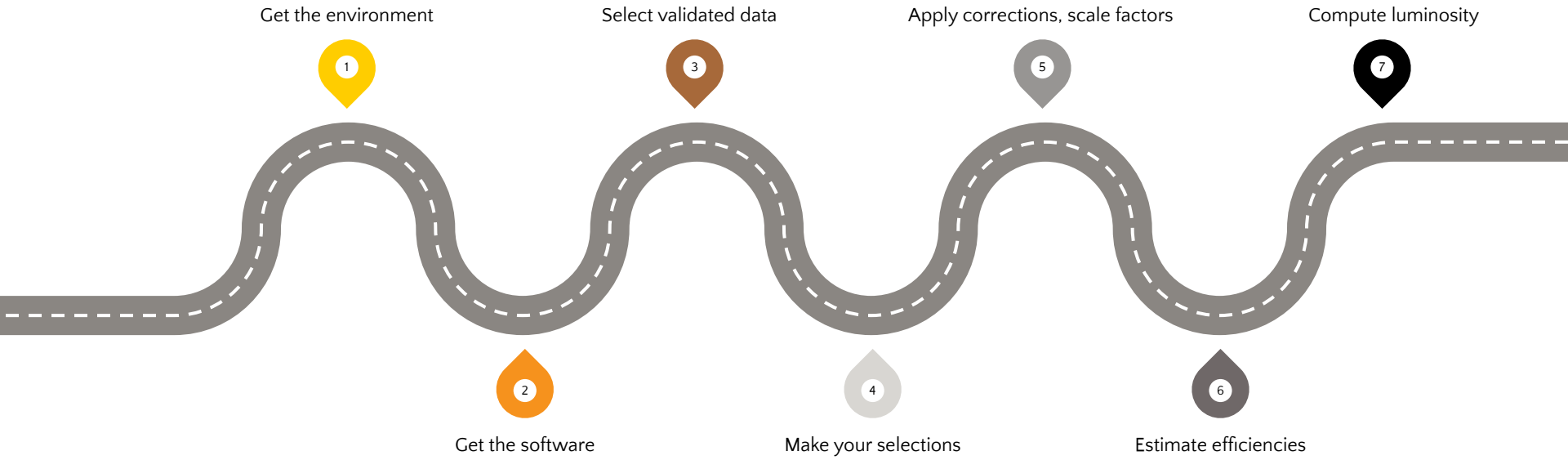
File Indexes

Filename	Size	
CMS_Run2015D_JetHT_MINIAOD_16Dec2015-v1_00000_file_index.txt	2.4 kB	List Files / Download
CMS_Run2015D_JetHT_MINIAOD_16Dec2015-v1_50000_file_index.txt	123.9 kB	List Files / Download

Contextual metadata - should cover a lot!



How to put this together?



Providing contextual metadata in a useful way

- Best through example workflows
  - automated, machine and human-readable
- Not easy to define (even with >1200 papers...)
  - partly because
    - analysis processes are complex
    - CMS data format supports a wide range of use cases
  - but also because we, as a community, have undervalued:
    - documentation
    - common tools
    - analysis code reuse.

3 Support to external analysts

- Open data portal information
- CMS Open data guide
- CERN open data forum
- CMS open data workshops

THE example code: POET

Physics Objects Extractor Tool

Put together by many people in the CMS open data group from various sources within CMS.

Covers Run1 AOD and Run2 MiniAOD

Not a negligible effort!

Yearly CMS Open data workshops

Next: 11-14 July!!!

4

Challenges

Open data have value only when in use but usability does not come for free

A little interlude to what it takes to open science to happen

4.1 Roles & responsibilities

Open science - what it takes to make it happen

Knowledge preservation
Data and software skills
Time

Infrastructure
Resources
Guidelines

Policies
Agreements
Practices

4.2

Best practices

To preserve the knowledge at the time of active analysis

Efforts are needed

Manage your code in versioned code repositories

Document everything from the start

Capture the steps in your analysis

Define easily reusable workflows

Package your analysis environment in software containers

Use continuous, automated testing

Best practices require time but they will pay off:

for the individual, for the group, and eventually, for open science!

5

Plans

and some topics for discussion...

Approximate plans

	2023				2024			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
2013 Heavy-ion data release preparations								
2023 CMS Open data workshop preparations								
2016 pp data release preparations (includes NanoAOD)								
2024 CMS open data workshop preparations (examples with NanoAOD)								
Update CMS open data guide								
2017 pp data release preparations (to be discussed and approved)								
Improve metadata for workflows								
Benchmarking workflows on public cloud resources								

Topics for discussion

Person-power

Not everyone agrees on open data

Usage patterns: our expectations vs reality

Value to OD efforts by big OD/OS projects?

Goal of open research data: research (by others), not outreach (by us)

Thank you!

Any questions ?

And thanks to SlidesCarnival for this free presentation template