CMS
Data and Analysis Preservation: Experience and plans

June 2, 2020
Software & Computing Round Table
BNL & Jefferson Lab

Hello!

I am Kati Lassila-Perini

- experimental particle physicist
- from Helsinki Institute of Physics (Finland)
- based at CERN
- coordinating data preservation and open access in the CMS experiment

✉ kati.lassila-perini @ cern.ch

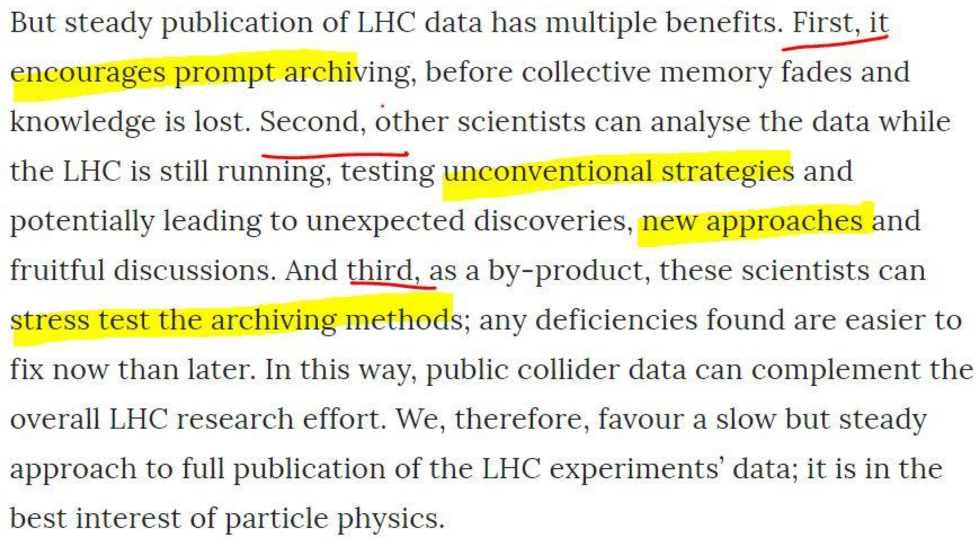
@KatiLassila

1.

Why open data?

Open data as a driving force to data and analysis preservation

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



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.

Use of CMS open data in research

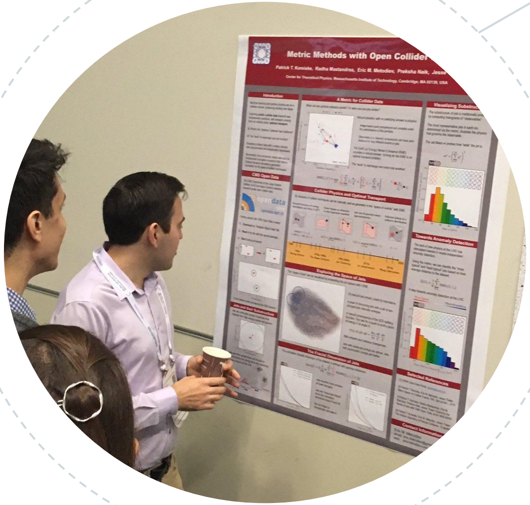
Examples:

- exploring experimental data with new methods
- jet and QCD studies
- machine learning studies
- root and other analysis tools development

Authors adhere to open science paradigm and

- share code e.g. [EnergyFlow python package]
- derived data, e.g. [Jet data in hdf5]

Research users have provided valuable and detailed feedback.


Metric Methods with Open Collider
Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, Jesse
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Introduction
A Metric for Collider Data
Visualizing Substruc

First [CMS Open data workshop] Sept 30 - Oct 2, 2020!

Use of CMS open data in education

Examples:

- CMS masterclasses
- Open data tutorials for teachers and students
- High-school teacher training
- Open data exercises in particle physics courses

All material in github free to use under a CC-BY.
Usable with free resources such as mybinder.org or google colab

Free research-level data can be **adapted** to various new use cases for education **without constraints**

Open data in use ➡ Data preservation

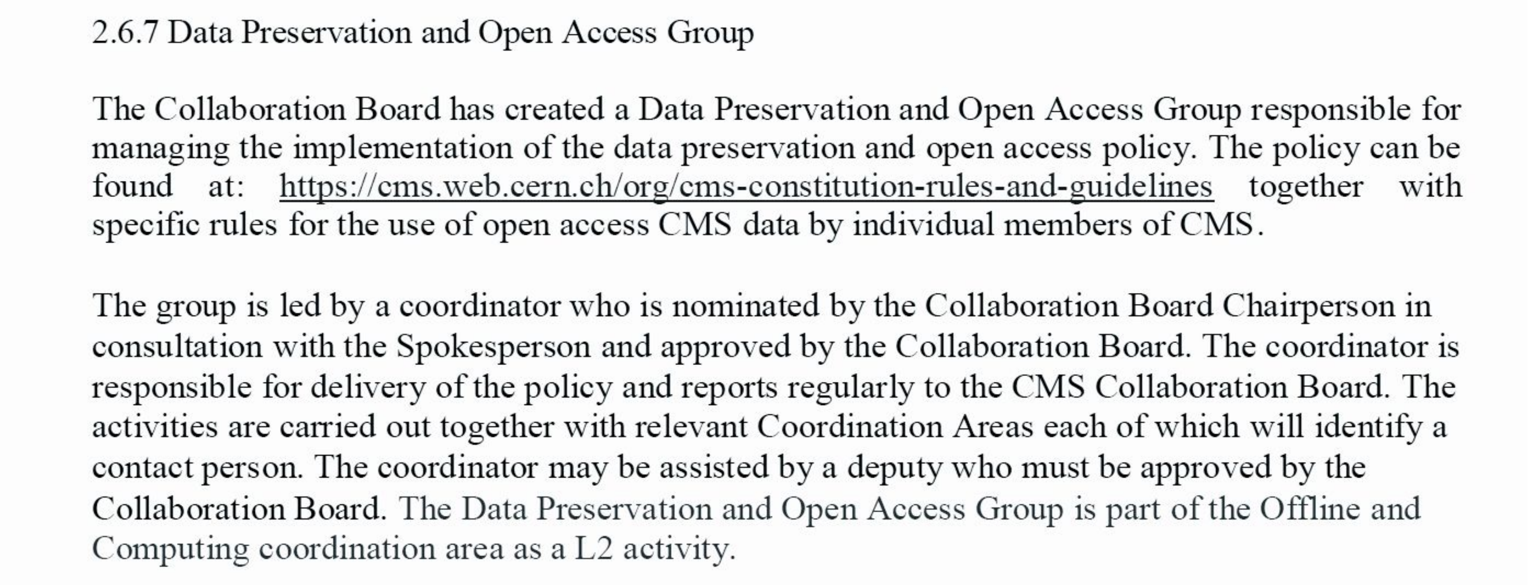
Open data makes data and analysis preservation to happen.

2.

What open data?

Open data policy and
Data preservation and open access (DPOA) group in CMS

CMS Data preservation and open access group mandate in CMS constitution:



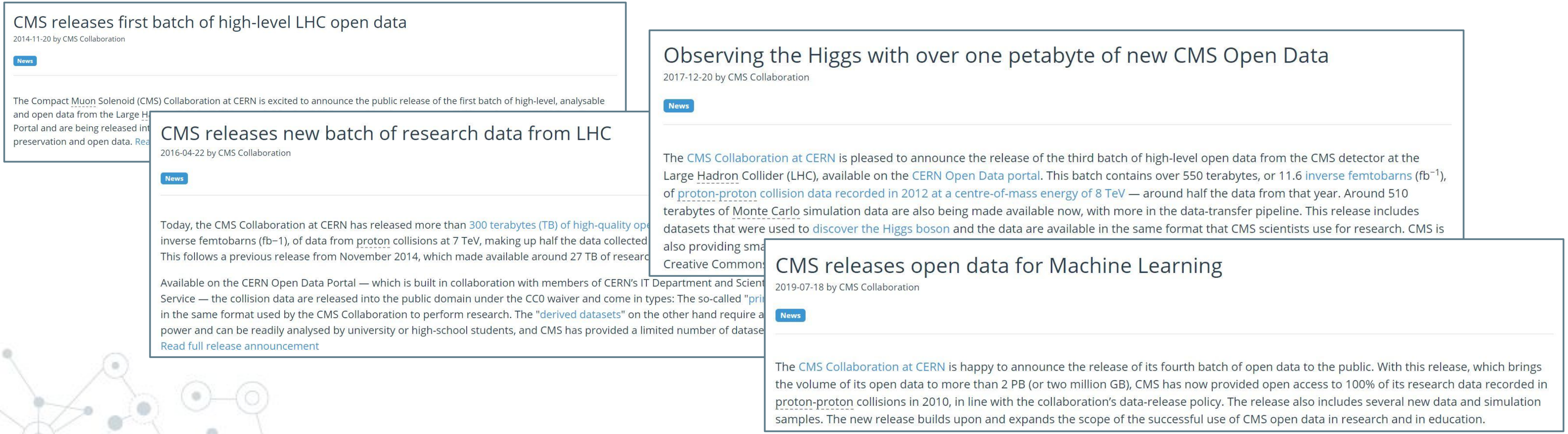
2.6.7 Data Preservation and Open Access Group

The Collaboration Board has created a Data Preservation and Open Access Group responsible for managing the implementation of the data preservation and open access policy. The policy can be found at: https://cms.web.cern.ch/org/cms-constitution-rules-and-guidelines together with specific rules for the use of open access CMS data by individual members of CMS .

The group is led by a coordinator who is nominated by the Collaboration Board Chairperson in consultation with the Spokesperson and approved by the Collaboration Board. The coordinator is responsible for delivery of the policy and reports regularly to the CMS Collaboration Board. The activities are carried out together with relevant Coordination Areas each of which will identify a contact person. The coordinator may be assisted by a deputy who must be approved by the Collaboration Board. The Data Preservation and Open Access Group is part of the Offline and Computing coordination area as a L2 activity.

CMS data preservation, re-use and open access policy

- The policy defines the CMS approach at different levels of data
- Approved in 2012, followed by the 1st data release in 2014, updated in 2018
- Is a statement of intentions put into practice through concrete actions:



CMS releases first batch of high-level LHC open data

2014-11-20 by CMS Collaboration

News

The Compact Muon Solenoid (CMS) Collaboration at CERN is excited to announce the public release of the first batch of high-level, analysable and open data from the Large H
Portal and are being released int
preservation and open data. Rea

CMS releases new batch of research data from LHC

2016-04-22 by CMS Collaboration

News

Today, the CMS Collaboration at CERN has released more than 300 terabytes (TB) of high-quality ope
inverse femtobarns (fb−1), of data from proton collisions at 7 TeV, making up half the data collected
This follows a previous release from November 2014, which made available around 27 TB of researc

Available on the CERN Open Data Portal — which is built in collaboration with members of CERN's IT Department and Scient
Service — the collision data are released into the public domain under the CC0 waiver and come in types: The so-called "pri
in the same format used by the CMS Collaboration to perform research. The "derived datasets" on the other hand require a
power and can be readily analysed by university or high-school students, and CMS has provided a limited number of datase
Read full release announcement

Observing the Higgs with over one petabyte of new CMS Open Data

2017-12-20 by CMS Collaboration

News

The CMS Collaboration at CERN is pleased to announce the release of the third batch of high-level open data from the CMS detector at the Large Hadron Collider (LHC), available on the CERN Open Data portal. This batch contains over 550 terabytes, or 11.6 inverse femtobarns (fb^{-1}), of proton-proton collision data recorded in 2012 at a centre-of-mass energy of 8 TeV — around half the data from that year. Around 510 terabytes of Monte Carlo simulation data are also being made available now, with more in the data-transfer pipeline. This release includes datasets that were used to discover the Higgs boson and the data are available in the same format that CMS scientists use for research. CMS is also providing sma
Creative Common

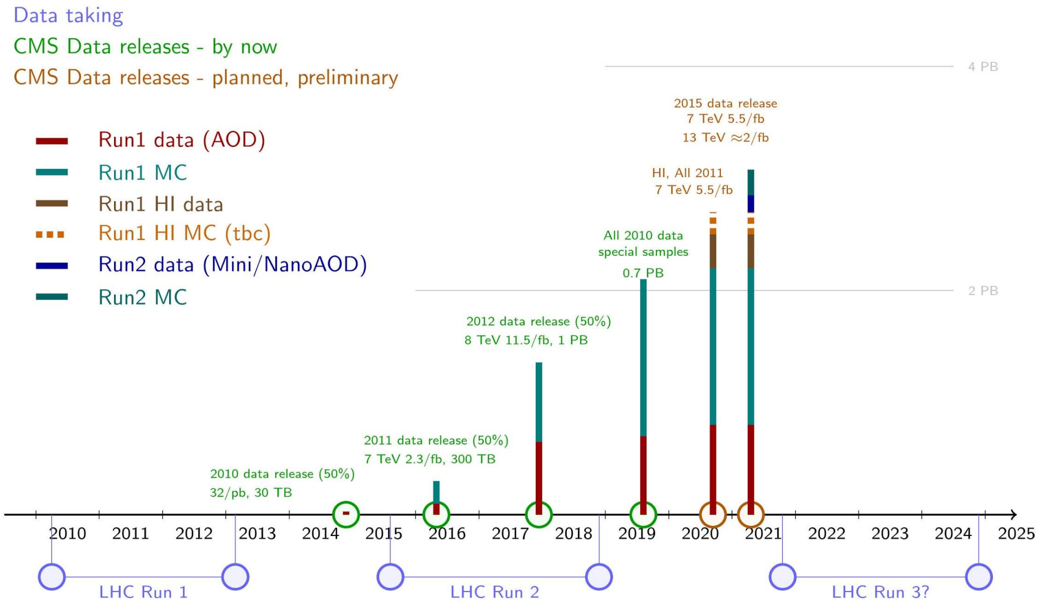
CMS releases open data for Machine Learning

2019-07-18 by CMS Collaboration

News

The CMS Collaboration at CERN is happy to announce the release of its fourth batch of open data to the public. With this release, which brings the volume of its open data to more than 2 PB (or two million GB), CMS has now provided open access to 100% of its research data recorded in proton-proton collisions in 2010, in line with the collaboration's data-release policy. The release also includes several new data and simulation samples. The new release builds upon and expands the scope of the successful use of CMS open data in research and in education.

Release schedule - data volumes



	Typical event size (kB)
Run1 (AOD)	**300**
Run2 (MiniAOD)	**32**
Run2 (NanoAOD)	**1**

Release data used in analysis ➡

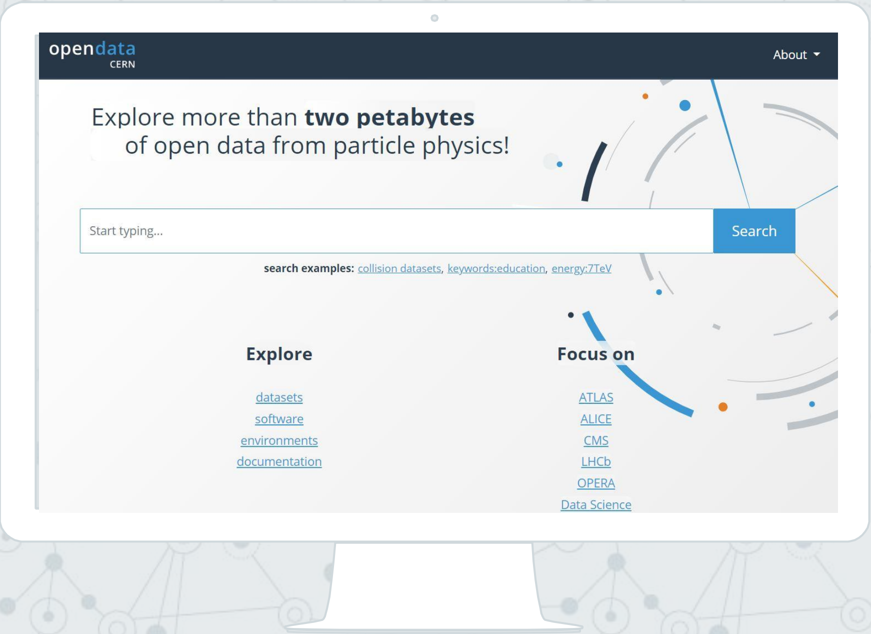
Validation done, rely on existing software and documentation

and: on reproducible analysis workflows (if you have them!)

3.

How - open data?

Release preparations and channels



Explore more than **two petabytes** of open data from particle physics!

Start typing...

search examples: collision datasets, keywords:education, energy:7TeV

Explore

datasets
software
environments
documentation

Focus on

ATLAS
ALICE
CMS
LHCb
OPERA
Data Science

CERN Open data portal

Serves the data, associated analysis artefacts, usage examples

Release preparations

Define legacy
  data and MC

CB approval
  release content and time

Transfer
  to OpenDataT3 - eospublic

Metadata
  from internal dbs

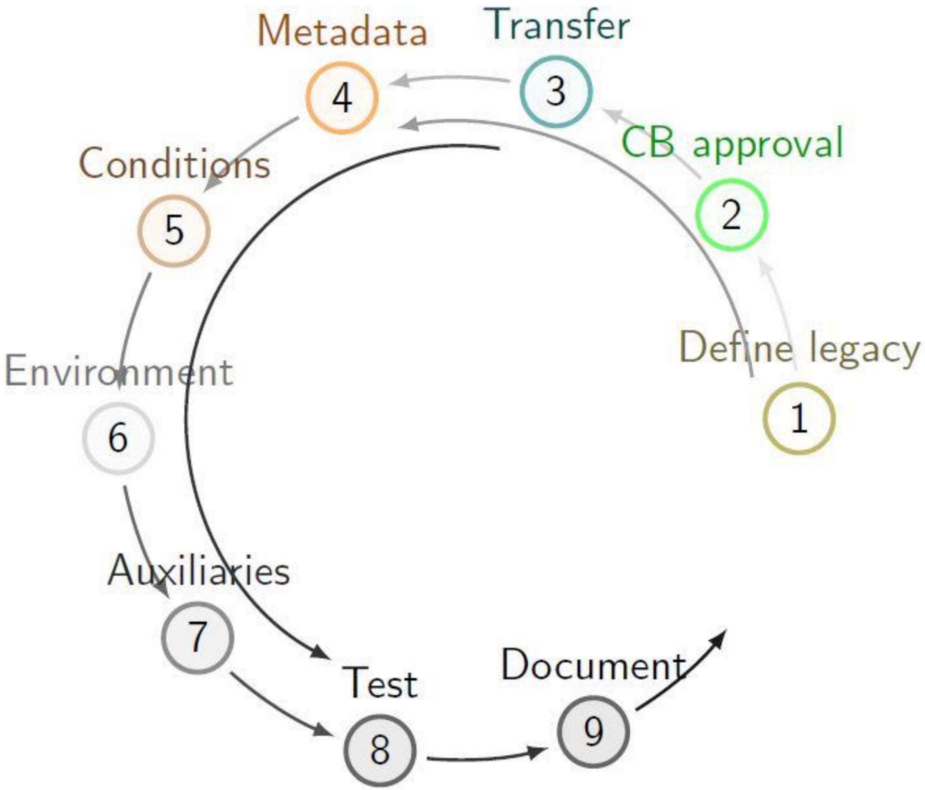
Conditions
  as sqlite files to cvmfs

Environment
  VM and container

Auxiliaries
  luminosity and good runs etc

Test
  on open data environment

Document
  prepare and update



Release preparations

Define legacy
  data and MC

CB approval
  release content and time

Transfer
  to OpenDataT3 - eospublic

Metadata
  from internal dbs

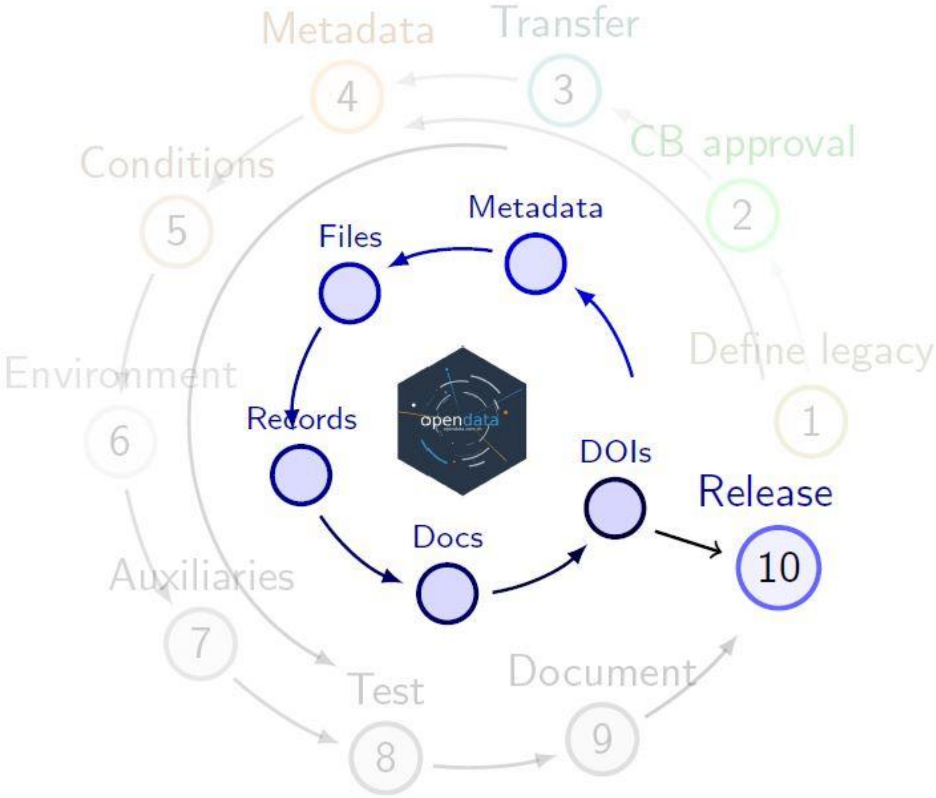
Conditions
  as sqlite files to cvmfs

Environment
  VM and container

Auxiliaries
  luminosity and good runs etc

Test
  on open data environment

Document
  prepare and update



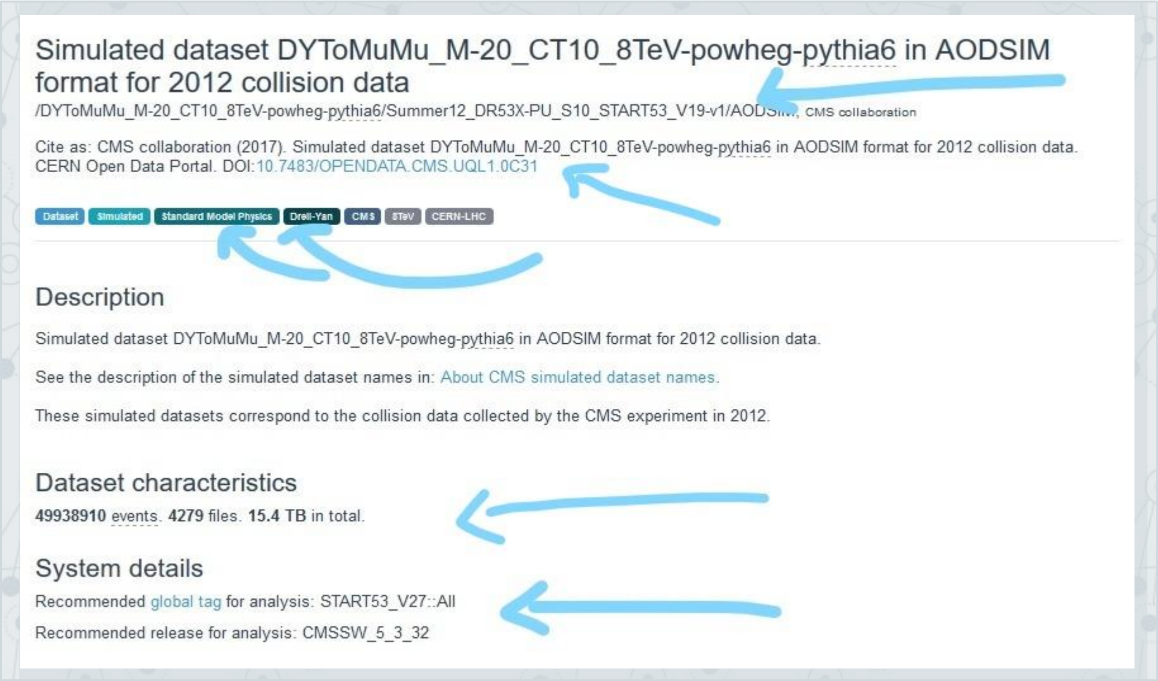
To get here, significant work by the CODP team

3.1

Provenance metadata

Data characteristics:
size/files/location
How these data were
acquired/generated/ reprocessed

An example record



Simulated dataset DYToMuMu_M-20_CT10_8TeV-powheg-pythia6 in AODSIM format for 2012 collision data

/DYToMuMu_M-20_CT10_8TeV-powheg-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM, CMS collaboration

Cite as: CMS collaboration (2017). Simulated dataset DYToMuMu_M-20_CT10_8TeV-powheg-pythia6 in AODSIM format for 2012 collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.UQL1.0C31

Description

Simulated dataset DYToMuMu_M-20_CT10_8TeV-powheg-pythia6 in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in: About CMS simulated dataset names.

These simulated datasets correspond to the collision data collected by the CMS experiment in 2012.

Dataset characteristics

49938910 events. **4279** files. **15.4** TB in total.

System details

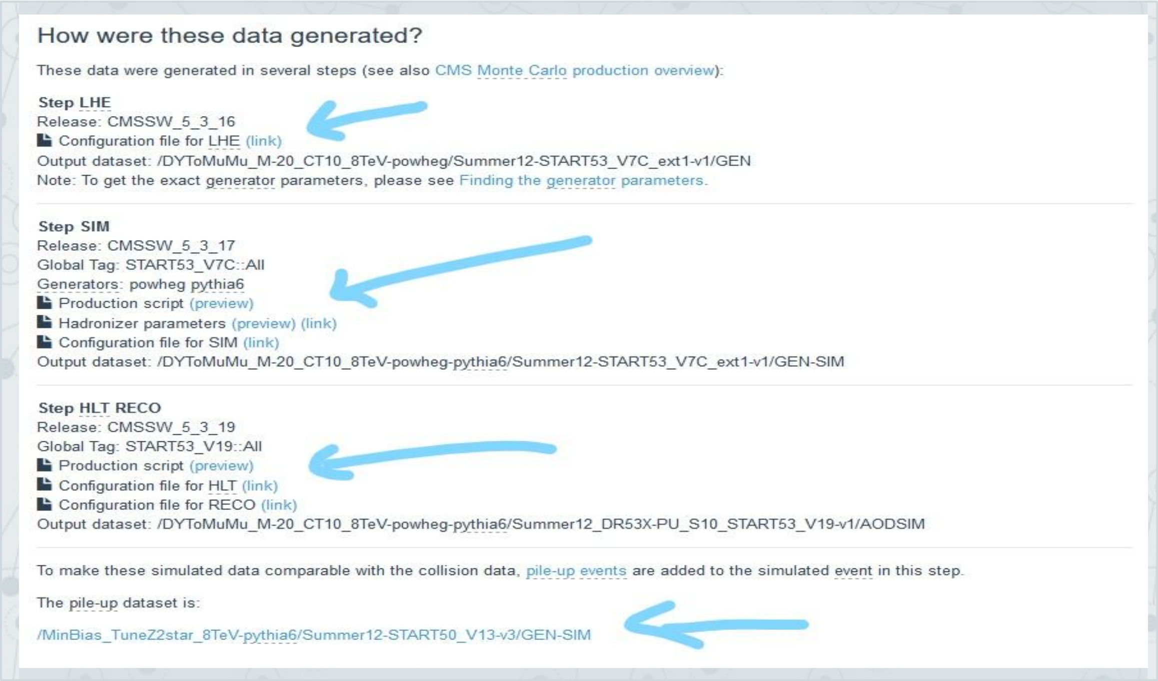
Recommended global tag for analysis: START53_V27::All

Recommended release for analysis: CMSSW_5_3_32

- Keep the original naming and file structure
- DOIs
- Tags for search

- Data characteristics (also for cross checking)
- Usage details

An example record



How were these data generated?

These data were generated in several steps (see also CMS Monte Carlo production overview):

Step LHE
Release: CMSSW_5_3_16
Configuration file for LHE (link)
Output dataset: /DYToMuMu_M-20_CT10_8TeV-powheg/Summer12-START53_V7C_ext1-v1/GEN
Note: To get the exact generator parameters, please see Finding the generator parameters.

Step SIM
Release: CMSSW_5_3_17
Global Tag: START53_V7C::All
Generators: powheg pythia6
Production script (preview)
Hadronizer parameters (preview) (link)
Configuration file for SIM (link)
Output dataset: /DYToMuMu_M-20_CT10_8TeV-powheg-pythia6/Summer12-START53_V7C_ext1-v1/GEN-SIM

Step HLT RECO
Release: CMSSW_5_3_19
Global Tag: START53_V19::All
Production script (preview)
Configuration file for HLT (link)
Configuration file for RECO (link)
Output dataset: /DYToMuMu_M-20_CT10_8TeV-powheg-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM

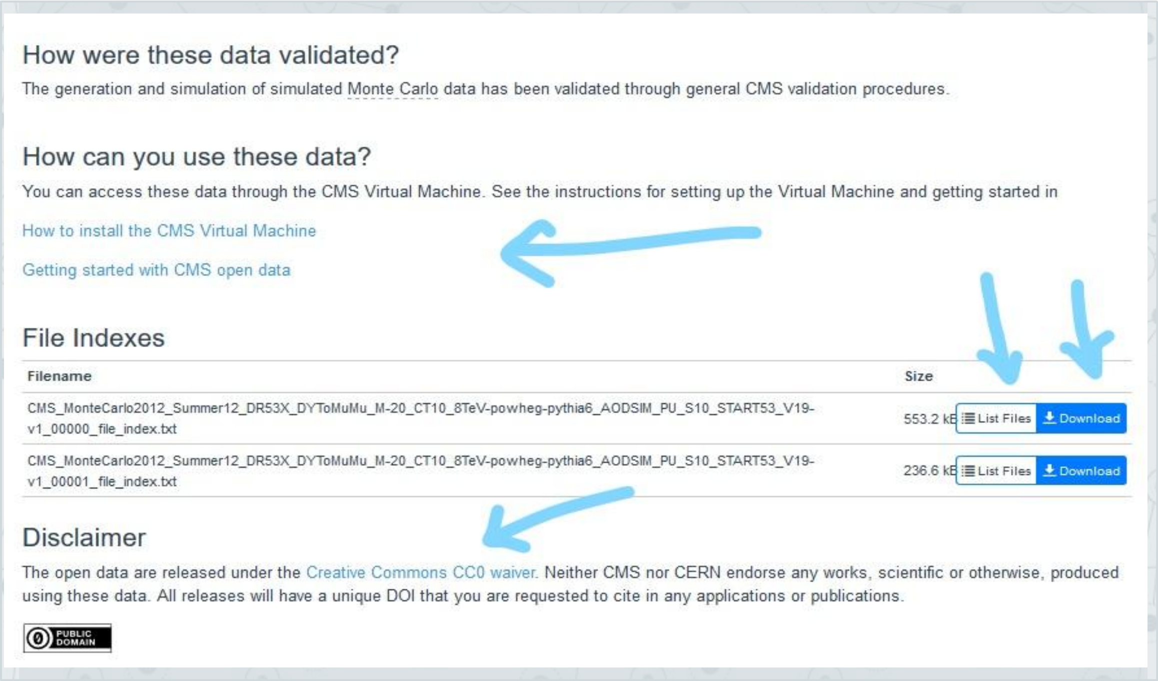
To make these simulated data comparable with the collision data, pile-up events are added to the simulated event in this step.

The pile-up dataset is:

/MinBias_TuneZ2star_8TeV-pythia6/Summer12-START50_V13-v3/GEN-SIM

- Full provenance for all production steps extracted from CMS-internal databases
  - Production release and conditions
  - Configuration files
- Pile-up information

An example record



How were these data validated?

The generation and simulation of simulated Monte Carlo data has been validated through general CMS validation procedures.

How can you use these data?

You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in

How to install the CMS Virtual Machine

Getting started with CMS open data

File Indexes

Filename	Size	
CMS_MonteCarlo2012_Summer12_DR53X_DYToMuMu_M-20_CT10_8TeV-powheg-pythia6_AODSIM_PU_S10_START53_V19-v1_00000_file_index.txt	553.2 kE	List Files Download
CMS_MonteCarlo2012_Summer12_DR53X_DYToMuMu_M-20_CT10_8TeV-powheg-pythia6_AODSIM_PU_S10_START53_V19-v1_00001_file_index.txt	236.6 kE	List Files Download

Disclaimer

The open data are released under the Creative Commons CC0 waiver. Neither CMS nor CERN endorse any works, scientific or otherwise, produced using these data. All releases will have a unique DOI that you are requested to cite in any applications or publications.

- Link to usage instructions
- Direct download file by file
- File index for xrootd access
- Creative Commons CC0

3.2

“Context” metadata

What additional artefacts are needed

How to use these data

Additional analysis artefacts

Validated runs

No special “open data” filtering, released data include all runs. A list of validated runs and lumi sections is provided.

Luminosity

Detailed luminosity tables and associated systematic uncertainty values are provided. Work ongoing for improved usability.

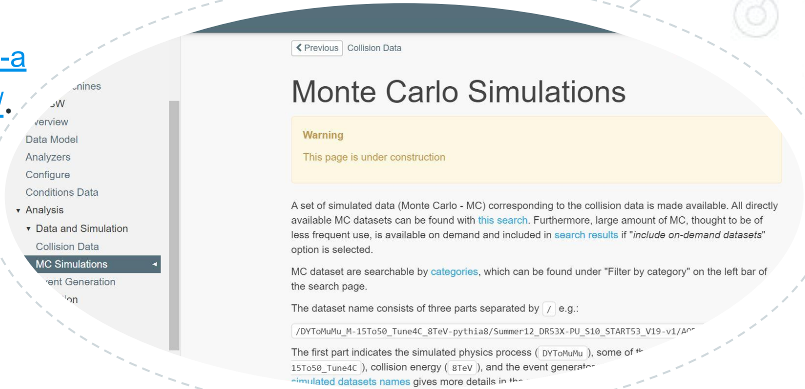
Condition data

Condition database snapshots for released data and MC are safeguarded in cvmfs. Good for analysis, and also for reconstruction from RAW and for MC generation.

OK, here are the data, now what?

- While access to data quick and easy, the learning curve to proper analysis is steep
  - Instructive reading from CMS OD users: https://profmattstrassler.com/2019/03/19/the-importance-and-challenges-of-open-data-at-the-large-hadron-collider/.
- Documentation is still scattered and incomplete
  - Working towards
    - CMS Open Data user guide
    - set of tutorial lessons

    for the CMS Open data workshop for theorists



Monte Carlo Simulations

Warning

This page is under construction

A set of simulated data (Monte Carlo - MC) corresponding to the collision data is made available. All directly available MC datasets can be found with this search. Furthermore, large amount of MC, thought to be of less frequent use, is available on demand and included in search results if "*include on-demand datasets*" option is selected.

MC dataset are searchable by categories, which can be found under "Filter by category" on the left bar of the search page.

The dataset name consists of three parts separated by `/` e.g.:

```
/DYToMuMu_M-15To50_Tune4C_8TeV-pythia8/Summer12_DR53X-PU_S10_START53_V19-v1/AO
```

The first part indicates the simulated physics process (`DYToMuMu`), some of th `15To50_Tune4C`), collision energy (`8TeV`), and the event generator simulated datasets names gives more details in th

Availability of **preserved analysis workflows** will be a game changer in the usability of open data

Collect all metadata as soon as possible

Systems change, knowledge fades out, recipes get lost, and people busy with data taking will not have time to help you with 10 y old data

4.

How to reuse?

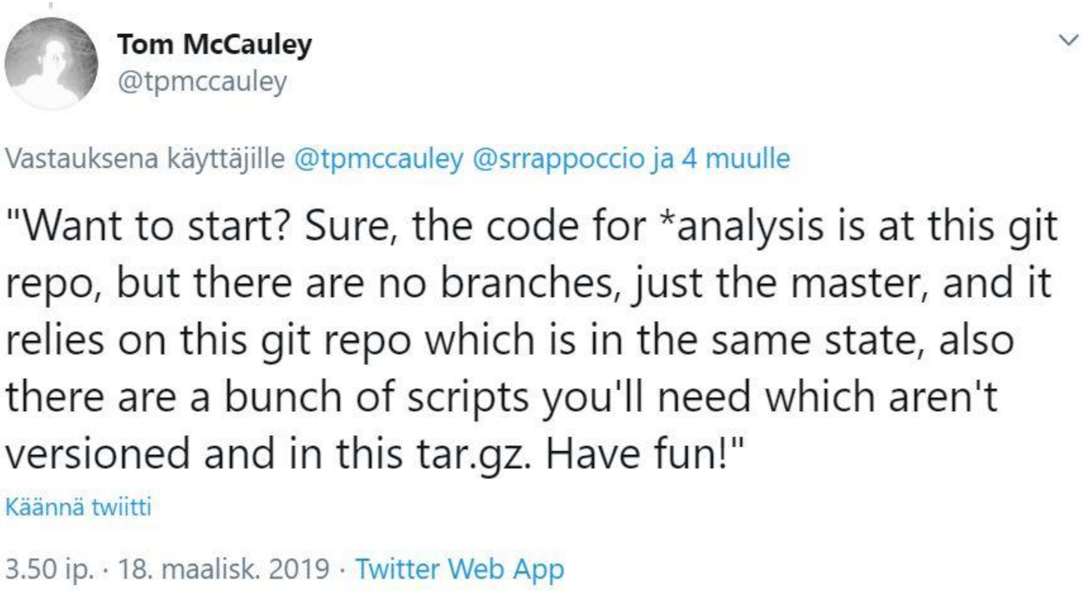
Preserving the knowledge

for internal benefit

(ongoing analyses)

for external use

(with open data)



Tom McCauley
@tpmccauley

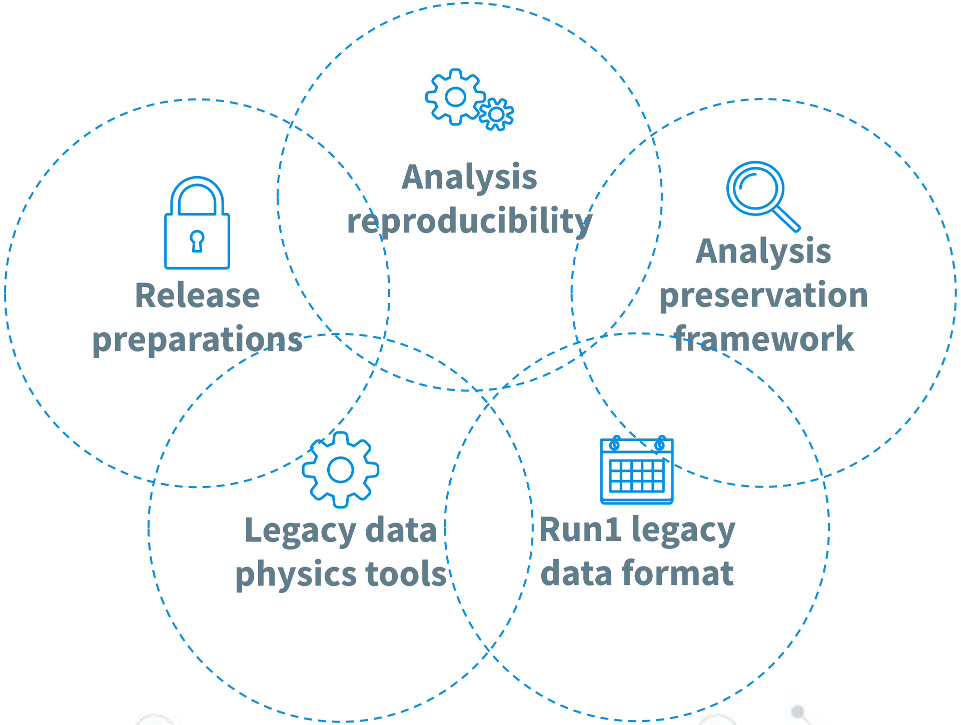
Vastauksena käyttäjille @tpmccauley @srrappoccio ja 4 muulle

"Want to start? Sure, the code for *analysis is at this git repo, but there are no branches, just the master, and it relies on this git repo which is in the same state, also there are a bunch of scripts you'll need which aren't versioned and in this tar.gz. Have fun!"

Käännä twiitti

3.50 ip. · 18. maalisk. 2019 · Twitter Web App

Not only open data - DPOA group activities:



Analysis reproducibility

Release preparations

Analysis preservation framework

Legacy data physics tools

Run1 legacy data format

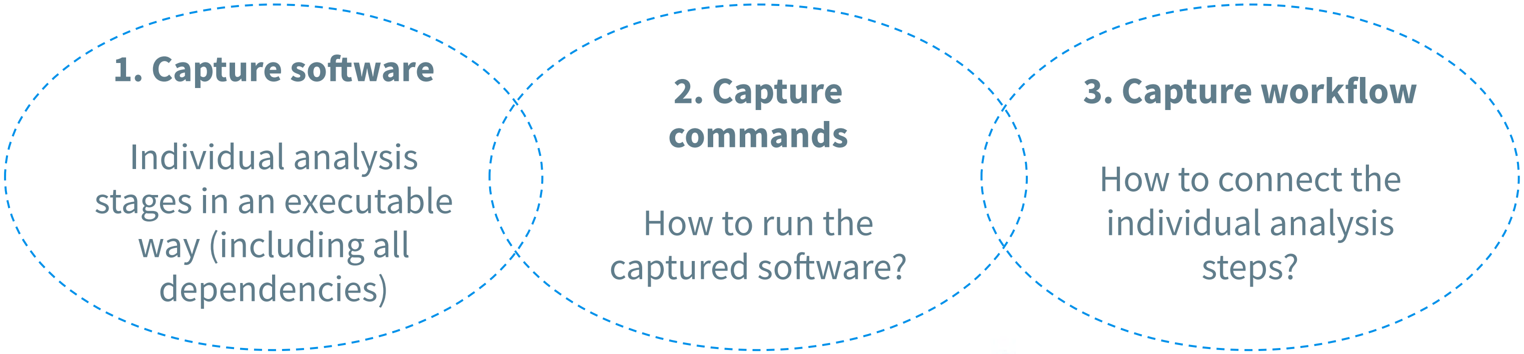
Increasing interest in analysis preservation

- Drafting of updated CERN Open Data Policy for the LHC experiments currently in progress
  - Explicit emphasis on preserved analysis workflows
- CMS policy covers it as well
  - “[..] Analysis procedures, workflows and code are preserved [..]”
  - we are starting to address the topic.
- Recent trainings received huge interest
  - ATLAS+CMS Analysis Preservation Bootcamp (February) oversubscribed by > factor 2
  - Virtual Pipelines training > 250 registrants (then closed registrations)

Analysts want to preserve their analyses, but they don’t know how!

Goal: Preserving the analysis implementation

- Goal: preserve analyses **during** the development/approval process already
  - Make this as **easy** as possible
- Continue **training** analysts
  - Teach use of tools (continuous integration, image building, …)
  - Develop additional tools where necessary (in particular for workflow automation)
- Aiming for a CMS-specific training as follow-up to the Virtual Pipelines training



1. Capture software

Individual analysis stages in an executable way (including all dependencies)

2. Capture commands

How to run the captured software?

3. Capture workflow

How to connect the individual analysis steps?

5.

Working together

Importance of common tools

Common data preservation services

“External” services

For limited-term projects, such as experiments, data preservation must rely on “external” service providers, not on the project itself.

With a close understanding

Very close understanding of data and their usage patterns are needed in order to provide these services.

Resources are limited, timeline is long:

avoid “single-experiment”, “single-person”, short-term solutions.

Our tools and services at CERN ♡

CODP

CERN Open data portal to store and serve the data and associated artefacts

ReANA

Data analysis platform for reusable and reproducible analysis workflows

CAP

CERN analysis preservation framework to catalogue and store analysis information

cvmfs

The CernVM File System to distribute the software

invenio

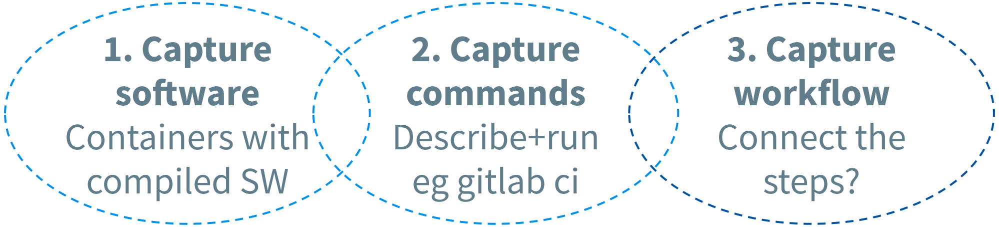
A library management software under CODP and many others

eos

Low-latency disk service to provide access to data through xrootd protocol

Call for ideas, discussion, collaboration

- We have experience and success in preserved, open data
  - We hope you can learn from our experience!
- Need and importance of preserved analysis workflows is obvious



1. Capture software
Containers with compiled SW

2. Capture commands
Describe+run eg gitlab ci

3. Capture workflow
Connect the steps?

  - 1 & 2: technically feasible, training and change of attitude needed
  - 3: human/machine readable workflow descriptions:

    available tools? do they match the needs of HEP computing? easy integration with analysis development workflow?

  - We hope we can learn from your experience!

Thanks!

Any questions?

Find me at:

kati .lassila-perini @ cern.ch

Credits

Thanks to my colleagues

- in the DPOA group in CMS
  - Edgar Carrera, Clemens Lange, 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
- in the CERN Open Data policy working group

Great thanks also to all CMS open data users!

And thanks to SlidesCarnival for this free presentation template