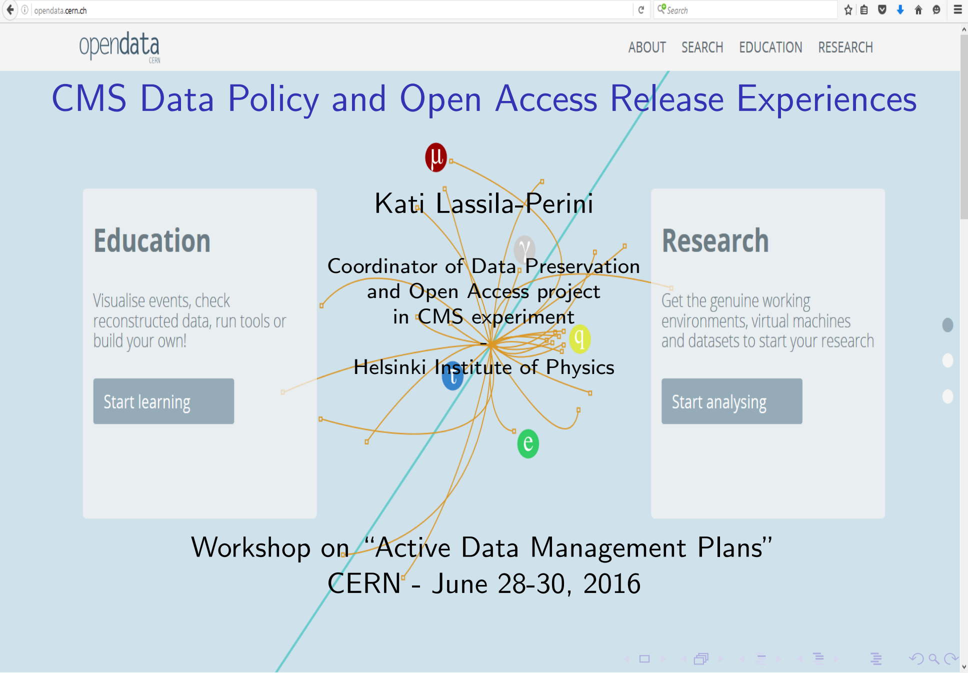
CMS Data Policy and Open Access Release Experiences



opendata.cern.ch

Search

opendata CERN

ABOUT SEARCH EDUCATION RESEARCH

Education

Visualise events, check reconstructed data, run tools or build your own!

Start learning

Research

Get the genuine working environments, virtual machines and datasets to start your research

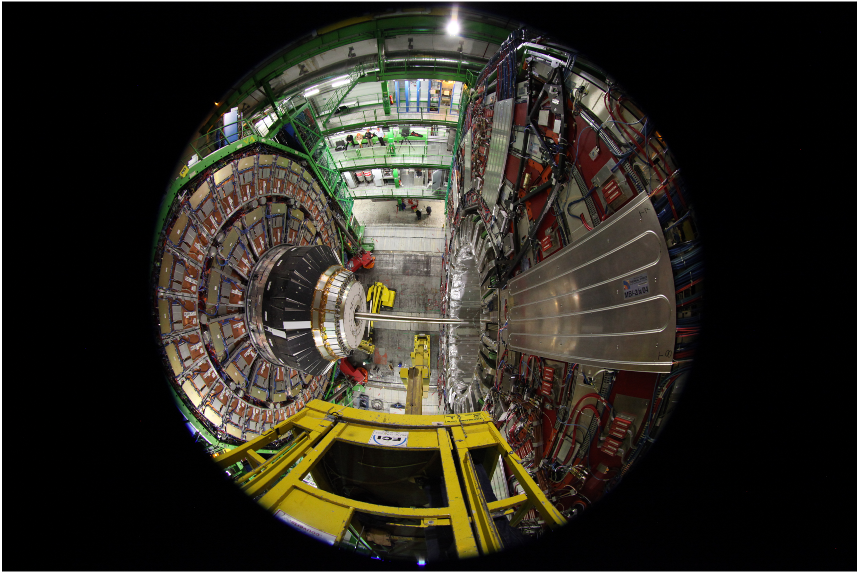
Start analysing

Kati Lassila-Perini

Coordinator of Data Preservation
and Open Access project
in CMS experiment
–
Helsinki Institute of Physics

Workshop on "Active Data Management Plans"
CERN - June 28-30, 2016

CMS experiment at LHC



CMS data levels and open data

- CMS has approved a data preservation, re-use and open access policy, which defines the approach to access to them at various levels:
  - Level 1 - Open access publication and additional numerical data
  - Level 2 - Simplified data for outreach and education
  - Level 3 - Reconstructed data and the software to analyze them
  - Level 4 - Raw data, and the software to reconstruct and analyze them.

CMS Open Data

- CMS continues publishing and promoting levels 1 & 2.
- CMS data releases at level 3 - reconstructed data:
  - November 2014: 28 TB of 2010 collision data
  - April 2016: $>$ 100 TB of 2011 collision data and $>$ 200 TB of simulated data.

▸ CMS data preservation, re-use and open access policy

The challenge: knowledge preservation

- In HEP, we are doing well with the “immediate” metadata, such as
  - beam conditions, event and run numbers, provenance information (raw data from which data have been reconstructed, the software version used in the reconstruction)...

  recorded together with the data records at the time of creation.
- We are doing poorly with the “context” metadata, such as
  - how to pick up the right objects in the data
  - how to know if there are additional selections, corrections...

  in general, the practical information needed to put the data in context and analyze them: information, which is readily available and even obvious at the time of the data analysis, but easily forgotten.

Open Data helps/forces us to meet this challenge

- Information must be collected and released together with the data.

CMS Open Data release

- Data
  - CMS collision data in format used in analysis by CMS physicists
  - For 2011 data, a partial set of simulated data included
    - ★ important for scientific use of data in HEP.
  - Run2 ($\geq$2015): less complete, more compact and easier-to-use format
- Tools
  - VM image of the computing environment
  - Access to the corresponding software and condition data
  - Access to data through streaming (xrootd) or direct download
- Instructions
  - Basic instructions to get started ($\approx$ 15 mins to setup) with examples
  - Basic description of the physics objects (data contents)
  - Some analysis examples
- Examples of derived datasets for different education and outreach contexts
  - Event display, online histogramming
  - Code to produce the derived datasets

What is on the portal: a walk-through

- Navigation

  Opendata.cern.ch entry page - Education/Research · Search functionlities

- Data

  Collision data records (primary datasets) with detailed data selection information

  Simulated data records with detailed production information

  Derived data records for easier use in different contexts

- Tools and instructions

  Setting up the VM computing environment · Instructions for getting started

- Validation benchmarks

  Validation utilities collection

- Examples

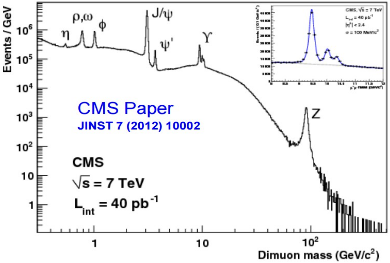
  Example code for a scientific study · Example code to extract simplified data

- Interfaces for outreach and education

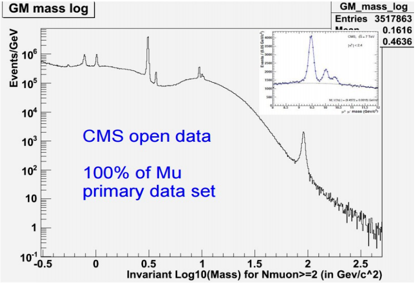
  Event display · Online histogramming examples

Open data benchmark/validation analyses

- Several benchmark analyses on for validation, and as examples:
  - high-level validation for each released primary dataset
  - feasible with the available data
  - possibility for comparison (later) with data at other beam energies
  - not too complicated but nevertheless interesting physics objects
  - published reference available.
- Act as a check list for re-usability!



Achim Geiser et al (DESY)



Examples of open data usage

- Scientific usage in HEP:
  - Ongoing analysis at MIT by a small group with a theorist, a post-doc and undergraduate
  - got started with the instructions on portal, and got help on volunteering basis from MIT and US CMS colleagues
  - aiming for a publication: Progress report
  - willing to contribute to the documentation to help other users.
- Scientific usage in other domains:
  - interest expressed for research into cloud computing security
  - machine learning
  - studies of statistical methods
- Teaching and training:
  - training of high-school teachers and undergrads http://coder.cern.ch
  - training of physicists CMSOpenData Exercise IFCA 2016
- External resources:
  - Computing resources https://cmsopendata.ifca.es/

Success metrics

- Do we support, for example, EU Data Principles?
  - Discoverable - readily found to exist by online search
    ▸ Search CMS experiment open data
  - Accessible - when discovered they can be interrogated
    ▸ A data record on the Open Data Portal
  - Intelligible - they can be understood
    ▸ High-level data description ▸ Further details
  - Assessable - the reliability of their source can be evaluated
    ▸ Details of the data processing step in the data record ▸ Validation benchmarks
  - Usable - they can be re-used
    ▸ Tools ▸ Instructions
- Work is ongoing, but we are on the right way to enable data to be exploited now and over the long term.
- Caveat:
  - What is now? What is long term?
  - Making data public does not make data simple.
  - Resources needed in addition to and beyond the operation and maintenance costs.

Outlook

- CMS experiment Data policy is a DMP (...short, lacks details...)
  - implemented and put in practice through the Open data release
  - will be complemented by "Analysis preservation", ongoing work to capture workflows and details connected to each publication.
- All this made possible by excellent collaboration with CERN services developing data preservation and open access tools
  - long-term DP: common, non experiment-specific solutions essential
  - great benefit from expertise in digital archiving and library services
  - see: Open Data and Data Analysis Preservation Services for LHC Experiments.
- Issues
  - data preservation must start when data analysis in ongoing, but we compete for resources with data taking, operation and new analyses.
  - detailed documentation lacking: only some reasonable thoughts...
  - data storage needs to be acknowledged and allocated.
- CMS has demonstrated that
  - complex data can be made usable
  - open data addresses the key issue of data preservation: knowledge preservation.