Implementing the data preservation and open access policy in CMS

K Lassila-Perini[1], G Alverson[2], I Cabrillo[3], A Calderon[3], D Colling[4], M Hildreth[5,6], A Huffman[4], T Lampén[1], P Luukka[1], J Marco[3], T McCauley[5,6] and L Sonnenschein[7]

[1] Helsinki Institute of Physics, Helsinki, Finland, [2] Northeastern Univ., Massachusetts, Boston, USA, [3] IFCA, CSIC-Univ. de Cantabria, Santander, Spain, [4] Department of Physics, Imperial College, UK, [5] University of Notre Dame, Notre Dame, IN, USA, [6] Fermi National Accelerator Laboratory, Batavia, IL, USA, [7] RWTH Aachen Univ., III. Physik. Inst. A,Aachen, Germany

on behalf of the CMS Collaboration

CMS policy

- CMS has approved a data preservation, re-use and open access policy[1], which defines the CMS approach to the preservation of the data and access to them at various levels of complexity.
- The implementation of the policy has been elevated to a dedicated project within the collaboration.
- CMS is looking for solutions, which could be usable for the other LHC experiments, and promotes common infrastructures.

Publications and additional data

- In parallel to the OA papers, publication of additional numerical data (easily extracted for further use) is encouraged, as well as initiatives such as RIVET[2] (easy comparison between data and MC event generators, and their validation).
  - Additional data has been added in HEPData to 52 publications
  - Rivet format has been provided for 21 analyses.
- For the distribution of these data, CMS relies on digital library services such as CDS, INSPIRE and, HEPData.

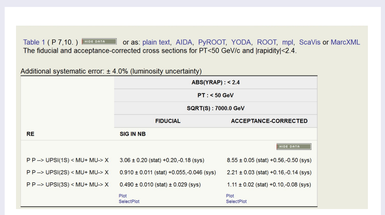


Table 1 (P 7,10.) or as: plain text, AIDA, PyROOT, YODA, ROOT, mpl, ScaVis or MarcXML
The fiducial and acceptance-corrected cross sections for PT<50 GeV/c and |rapidity|<2.4.

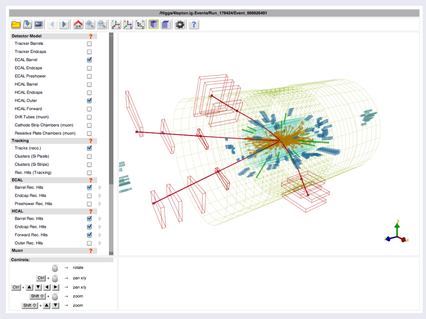
Additional systematic error: ± 4.0% (luminosity uncertainty)

	ABS(YRAP) : < 2.4	
	PT : < 50 GeV	
	SQRT(S) : 7000.0 GeV	
	FIDUCIAL	ACCEPTANCE-CORRECTED
RE	SIG IN NB	
P P --> UPSI(1S) < MU+ MU-> X	3.06 ± 0.20 (stat) +0.20,-0.18 (sys)	8.55 ± 0.05 (stat) +0.56,-0.50 (sys)
P P --> UPSI(2S) < MU+ MU-> X	0.910 ± 0.011 (stat) +0.055,-0.046 (sys)	2.21 ± 0.03 (stat) +0.16,-0.14 (sys)
P P --> UPSI(3S) < MU+ MU-> X	0.490 ± 0.010 (stat) ± 0.029 (sys)	1.11 ± 0.02 (stat) +0.10,-0.08 (sys)

Example of an additional data table attached to a publication.

Data for outreach and education

- Since 2010, several data sets released in a simplified format.
  - Used in the masterclasses developed and organized by QuarkNet[3] and the International Particle Physics Outreach Group (IPPOG)[4].
- Various tools are available for examination of the events, e.g. an online event display.


A Higgs candidate event shown with the online event display.

Legacy data sets

- No raw data will be deleted.
- All 2011-2012 collision and simulated data are being reprocessed into a legacy data set with a single CMS Software version.
- Duplicates of reprocessed data sets are regularly deprecated.

Bit-level data preservation

- At the bit-level, CMS computing model offers a solid base for long-term data preservation.
- Looking forward to the program of work of the bit-level data preservation working group under HEPiX forum, within WLCG.

Analysis preservation

- The biggest challenge: the loss of knowledge and expertise.
- CMS does very well in recording the "immediate" metadata:
  - event and run numbers, beam conditions, software versions used in the data reprocessing...

  but poorly on the "context" metadata,
  - i.e. practical information needed to put the data in context and analyze them.
- CMS is actively looking for solutions to enable easy recording of the detailed knowledge, readily available at the time of the active physics analysis, but quickly forgotten.
- The Invenio team at CERN, with input from CMS, will setup a prototype of a tool to make recording and documenting of the workflow and intermediate data easy.
- The DASPOS project[5] is studying the problematics of the reproducibility of physics results in a wider cross-experiment and cross-disciplinary context.

Long-term validation

- The goal of the long-term validation project is to extract and record the necessary information and tools in order to be able to validate the data and software in case of re-use in the longer-term future:
- A set of reference plots being defined covering physics results: from individual objects (muons, electrons,...) to high level physics signatures involving basic analysis selections.
- Using the CMS powerful tools for validation of software releases and for data quality monitoring.

Open access

- Preparing for a public release of part of 2010 collision and simulated data in AOD format, appropriate for physics analysis, and accompanied by stable, open source software needed for a number of example analysis and suitable documentation.
- A virtual machine (VM) image has been prepared, in a format usable by the freely-available VirtualBox application.
- For the access to the data, the initial work flow is kept as close to the standard one as possible, which uses xroot. An xrootd server has been commissioned with anonymous read-only access, further limited by firewall to include only those sites involved in the testing phase.
- Preparing high-school level classroom applications chosen as a pilot use-case[6] - tools to process the data and documentation into an appropriate format, studies on:
  - a common definition of data contents for most HEP experiments
  - a common HEP ontology with Linked Data methods.

Outlook

- Data preservation must start when the data are created - CMS aims to integrate the data preservation approach in day-to-day work.
- CMS seeks for generic solutions, making possible the use of common approaches with other experiments and disciplines.
- CMS works in collaboration with DPHEP, the CERN-IT department and other laboratories to define different tools and services for data preservation and open access.

References

[1] CMS Collaboration, "CMS data preservation, re-use and open access policy": https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=6032
[2] http://rivet.hepforge.org/
[3] http://quarknet.fnal.gov/
[4] http://ippog.web.cern.ch/
[5] https://daspos.crc.nd.edu/
[6] Open CMS Data Finland: https://twiki.cern.ch/twiki/bin/view/HIPCMSExperiment/CMSOpenDataProject