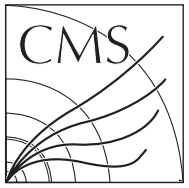
Available on CMS information server

CMS CR -2012/065



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

10 May 2012 (v3, 22 May 2012)

Preparing for long-term data preservation and access in CMS

Kati Lassila-Perini for the CMS Collaboration

Abstract

The data collected by the LHC experiments are unique and present an opportunity and a challenge for a long-term preservation and re-use. The CMS experiment has defined a policy for the data preservation and access to its data and is starting its implementation. This note describes the driving principles of the policy and summarises the actions and activities which are planned in the starting phase of the project.

Presented at *CHEP 2012: International Conference on Computing in High Energy and Nuclear Physics*

Preparing for long-term data preservation and access in CMS

Kati Lassila-Perini[1], David Colling[2], Sudhir Malik[3], Jesus Marco[4], Elizabeth Sexton-Kennedy[5], Lucas Taylor[5], Roberto Tenchini[6] for the CMS Collaboration
[1]Helsinki Institute of Physics, Helsinki, Finland
[2]Imperial College, London, UK
[3]Fermilab/University of Nebraska, USA
[4]Instituto de Fisica de Cantabria (IFCA, CSIC-UC), Santander, Spain
[5]Fermilab, PO Box 500, Batavia, IL 60510-5011, USA
[6]Istituto Nazionale di Fisica Nucleare, Pisa, Italy

E-mail: kati.lassila-perini@cern.ch

Abstract. The data collected by the LHC experiments are unique and present an opportunity and a challenge for a long-term preservation and re-use. The CMS experiment has defined a policy for the data preservation and access to its data and is starting its implementation. This note describes the driving principles of the policy and summarises the actions and activities which are planned in the starting phase of the project.

1. Introduction

The long-term preservation of the data and access to them is gaining increasing attention in the high energy physics (HEP) experiments. The data are unique and considerable resources have been used in their collection, and there may be a large scientific benefit in keeping these data accessible and re-usable. The HEP experiments having ended the data-taking or in the final analysis phase have discussed the concerns and practices connected to the data preservation in the context of Data Preservation in HEP (DPHEP) working group[1]. The experiments at the LHC have recently joined this effort and are preparing to face the issue in a timely manner. The Compact Muon Solenoid (CMS) collaboration[2] has prepared a policy, which defines the data preservation practices and open access principles.

The key question and challenge for the data preservation in CMS is whether it will be possible for CMS collaborators to re-analyse the current CMS data in few years in the future. The LHC data is very rich and there are many potential use-cases for re-use such as cross-checks with different centre-of-mass energies, luminosities or pile-up conditions, and lower trigger thresholds. In short term, re-analysing the data is feasible and indeed is already being done in CMS for the 2010 data. However, it needs to be assured with proper planning and resources that it will be possible to do so also in long term.

Besides the internal interest of CMS to preserve its data, there is a growing interest in the public domain in having access to these data, which is reasonable, considering that the research is publicly funded. For these reasons, the CMS Collaboration Board (CB) mandated a working group to prepare a policy on data preservation, re-use and open access. The policy[3] was endorsed by the CB in March 2012, and CMS will now start implementing the measures that will guarantee the data availability in long term. This note summarises the policy and first steps foreseen for its implementation.

2. Guiding principles in preparing the policy

The financial and human resources of the CMS collaboration are built up to fulfill its main goal: build and operate equipment to study the fundamental properties of matter with the data produced at the LHC. The organization of the collaboration and the current practices are formed to best achieve this goal. Therefore, for the most efficient use of the available resources, the following three principles were used as a guidance in preparing the policy on data preservation, re-use and open access.

- The measures concerning data preservation and access should be in line with the current CMS practices and these measures should benefit the CMS collaborators in their daily work.
- All data preservation activities should be a natural long-term extension of the current practices.
- The data policy was required to include a protection of the collaboration from external use cases concerning open access that could generate burdens beyond the available funding and resources.

The CMS data policy is a commitment from the collaboration to preserve data and to give public access to part of it. All technical details will be addressed in the implementation phase, which is now only starting.

3. The policy

The data preservation and open access policy reflects the CMS collaboration's concept of openness and protection, and it addresses the requests on this topic to CMS from the funding agencies. The key message of the policy is that CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. To that extent, CMS will provide open access to its data after a suitable embargo period, which allows CMS collaborators to fully understand their scientific potential. As illustrated in figure 1, the policy describes the CMS approach to the questions: What is the data to be preserved? Who will be able to access the data and when will the data be available? In which terms the open data usage can be used?

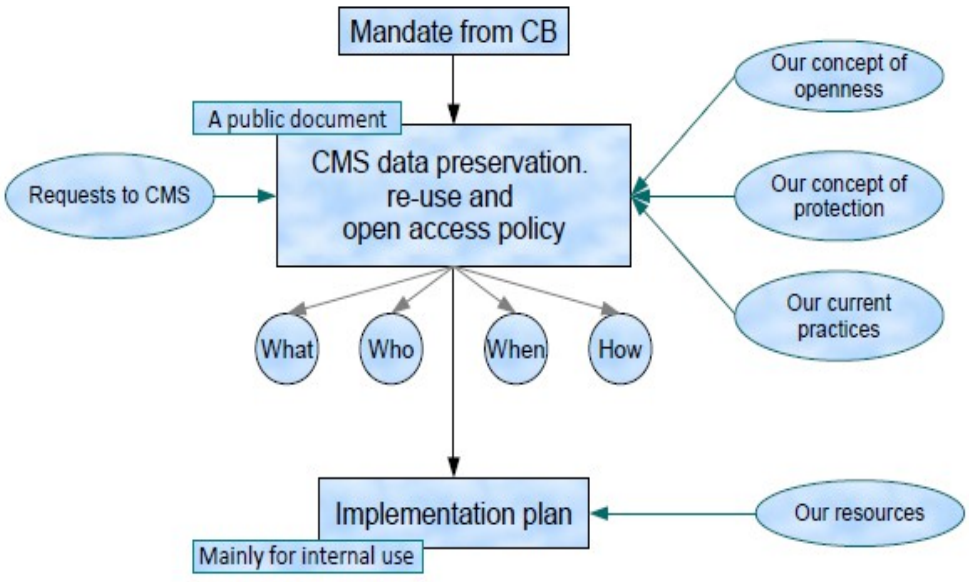


Figure 1. Preparing the CMS data preservation, re-use and open access policy.

3.1 What is data?

The CMS policy on data preservation and access defines the different levels of data using a four-level description common in HEP[4]:

- Level 1: Publications, additional documentation to put the results in context and understand the analyses procedures, some additional numerical data which did not or could not appear in the publications (e.g. cross sections as a function of multiple variables, data behind figures).
- Level 2: Simplified data formats (e.g. multi-dimensional distributions of analysis variables, four-vectors of particles/jets, energy clusters and tracks) for several levels of immediate re-use: theory interpretations, limited analyses, education, outreach.
- Level 3: Reconstructed data and simulations, together with the software, analysis workflows and documentation needed to access the data, understand them, reproduce published analyses, perform new analyses not requiring re-reconstruction of the data or new simulations.
- Level 4: Raw data and the software and documentation needed to access, reconstruct and analyse them.

3.1 Who will access the data?

The CMS collaborators have immediate access to the CMS data, and the membership conditions of the CMS collaboration are fully defined in the CMS constitution. The role of the policy on data preservation, re-use and open access is to define the terms in which the CMS data will be made available outside the collaboration. Possible re-use of the data may be done by collaboration members long after the data are taken, experimental and theoretical HEP scientists who were not members of the collaboration, educational and outreach initiatives, and citizen scientists in the general public. The policy, however, does not make any distinction of these different user profiles and the data, which will be made available for open access, will be available to everyone.

3.3 When will data be public?

CMS will provide open access to its data with appropriate delays depending of the type of the data.

- At level 1, the additional data complementing a CMS paper or note are made available at the moment of the publication.
- At level 2, simplified data format samples are released promptly and accepted for a public release case by case by the Collaboration Board.
- At level 3, public data releases, accompanied by stable, open source, software and suitable documentation, will take place yearly during long LHC machine shut-downs and at best efforts during running periods. During the lifetime of CMS an upper limit on the amount of publicly available data, compared to those still only available to the collaboration, will be up to 50% of the integrated luminosity collected by CMS. Data will be usually released 3 years after data taking, but the Collaboration Board could, in exceptional circumstances, decide to release some particular data sets either earlier or later.

The raw data formats of level 4 are not useful for analysis and will not be included in the public data release. Only level 3 formats after final calibration and reconstruction will be made available. These are the data formats used also internally in CMS for analysis.

3.4 In which terms can the open data be used?

The data will be released under the emerging standard Creative Commons CC0 waiver5. Data will also be identified with persistent data identifiers, and it is expected that the third parties cite the public CMS data through these identifiers, so that its re-use can be monitored and contribute to the assessment of the impact of the LHC program.

3.5 The first public release as a stress-test exercise

A first data release is planned to happen for the low-energy and low-luminosity runs of the LHC by 2013, as a stress-test exercise of the entire preservation, re-use and access chain. This release will be followed by a full analysis of the procedure and the experience will be evaluated by the Collaboration

Board in 2014 and in absence of unexpected overhead to the Collaboration the public data release will be accepted as a standard practice.

4. Current CMS practices

While preparing the policy document, the CMS way of operation was reviewed for each level of data complexity in order to identify possible bottle-necks for long-term data preservation and re-use. In general, it was found that the CMS way of operation is well aligned with a data preservation in long term. Furthermore, open access is already provided at levels 1 and 2.

4.1 Level 1 – Publications and additional data to them

The CMS physics results are published and stored at external, persistent open-access repositories[5]. The CMS Collaboration Board explicitly encourages the publication of numerical data used for relevant figures (plots, histograms, etc.) of CMS papers published in peer-reviewed journals. This policy complements the "open access" publication of articles, aimed at reaching out to the largest possible community without financial barriers.

This practise was already in place by the time of CMS' first publication on collision data[6], where the numerical data were added to the public journal web page as supplementary information. This example has been followed for subsequent papers, whenever useful. All journals in which CMS publishes allow the presence of ancillary information.

Numerical data related to CMS papers are also regularly sent to the HEPDATA database[7] in Durham. The HEPDATA system is well structured, well maintained and has a long tradition of collecting important High Energy Physics data. The HEPDATA group traditionally collects and stores in its database information related to standard model measurements, such as cross-sections, structure functions, events shapes. Recently, CMS and HEPDATA have started to develop ways to store information for searches, in form of confidence levels or by other means convenient for data interpretation. The policy is to send to HEPDATA "corrected information", which can be used by physicists outside the experiment without the need of a full simulation of the CMS detector. It is expected that INSPIRE[8] will provide a link between the publication and these additional data sets. Possibly, INSPIRE will also make these numerical data citable.

4.2 Level 2 – Simplified data-formats

Simplified data-formats can be used for high-level analysis tasks beyond the numerical information described at Level 1 and for educational and outreach purpose. CMS has already made small samples of selected interesting events available to educational programs targeting high school students. The text based .ig file format is human-readable and largely self-explanatory. The files are standard JSON (JavaScript Object Notation) files which means that they can easily be read using C++, Python or other programming languages.

CMS is developing software, jointly with the Quarknet and I2U2 programmes, that is being used in IPPOG Masterclass and I2U2 web-based e-lab educational exercises. These enable students to examine individual events from .ig files in a web-based iSpy event display program, to learn how different particles are detected and to classify decays (e.g. leptonic decays of J/psi, Upsilon, W, Z, etc.). They also perform simple statistical analyses using the same .ig format event samples.

The CMS Collaboration Board has explicitly approved all requests to make event samples public. So far the following have been approved:

- December 2010: approval for 2000 J/psi → mu mu, 2000 Upsilon → mu mu, 500 Z → mu mu and 500 Z → ee.
- April 2011: approval for 2000 J/psi → ee, 2000 Upsilon → ee, 1000 W → e nu, 1000 W → mu nu, 100k di-electron events, 100k di-muon events, 100k di-jet events all in the invariant mass range 2 – 100 GeV
- April 2012: approval for the first 50 pb^{-1} of single muon triggered primary datasets of 2010 for a tt analysis and the corresponding simulated background samples in a 4-vector format

All samples, from the above lists, that have already been selected, converted to .ig format and made public[9].

4.3 Level 3 – Reconstructed data and the software and documentation to access and analyse them
The reconstructed data are kept usable by forward-porting them to a current release of CMS Software (CMSSW). This reprocessing occurs when necessary and is a regular part of making the data available to the CMS collaboration. It requires that the calibration, conditions and trigger data for these datasets are kept. The reprocessing of the Monte Carlo samples does not happen systematically but on request by the physics groups.

The analysis software is stored in the common CMSSW cvs repository. The analysis software consists of common physics analysis tools which are a validated as part of each release. Physics group specific shared code is stored in group-specific cvs directories and analysis-specific user code in the a repository which is not a part of the CMSSW release.

The basic documentation and instructions on how to perform an analysis are given in the CMS WorkBook[10,11]. The WorkBook is regularly reviewed and updated. More detailed information is recorded in the CMS SWGuide[10,12]. Regular tutorial sessions are organized on the use of the physics analysis tools and a collection of well organized and updated tutorial material has helped to establish a common analysis tool framework, which in turn, helps to expedite the analysis and its approval by CMS[13].

The details of each analysis are recorded in internal Analysis Notes. The analysis guidelines require that standard procedures should be used in the analysis and that the software used in the analysis should be reviewable. These requirements are not currently enforced. The information necessary to review the analysis during the approval phase is linked through a CMS-specific analysis database. The full technical details in the analysis workflow are, however, not recorded.

4.4 Level 4 – Raw data and the software and documentation to access, reconstruct and analyse them
The CMS makes two 'custodial' copies of all raw event data[14]. The first is stored at the CERN Tier-0 centre, and the second distributed between Tier-1 centres. Neither of these copies are regarded as purely as a backup copy, and both will be used for reprocessing when required in order to optimise the overall efficiency of the system. By accepting a fraction of the raw data a Tier-1 centre is explicitly undertaking the long term stewardship of those data. This involves ensuring that the underlying mass storage system is protected against controllable risks such as hardware or media failures, environmental hazards, or breaches of security.

Each Tier-1 Centre forms an integral part of the central data handling service of the LHC Experiments[15]. It is thus essential that each such centre undertakes to provide its services on a long-term basis (initially at least 5 years). Also that each Tier 1 centre makes its best efforts to upgrade its installation and services in line with the expected growth of LHC data volumes and analysis activities. Specifically each centre records on archival permanent mass storage its fraction of the processed real and simulated data.

The simulation and reconstruction software is stored in the common CMSSW cvs repository. The software is validated for each release and it is required that each new release maintains the ability to read properly validated RAW data files written with older CMSSW releases.

The documentation of the simulation and reconstruction software is partly covered in the CMS WorkBook and the SWGuide. However, at the moment, these documents may not systematically cover all aspects of the production chain. The software packages and the classes therein stored in the CMSSW repository are only partially documented in the the CMSSW Reference manual, which in principle offers a version-dependent semi-automated documentation system for all material in the CMSSW repository.

5. Implementation plan

The policy on data preservation, re-use and open access has been approved by the CMS CB in March 2012 and the implementation phase is only starting. The implementation of the policy will be led by a project coordinator who is responsible for managing the proposed activities and reporting to the CB. The project coordinator has a duty to inform and manage the various stakeholders and is responsible for delivery of the policy.

It is envisaged that in short-term, within a year of the approval of the Policy, current practices in providing additional data (Level 1) and distributing simplified data-samples for outreach (Level 2) should be promoted and consolidated and become a routine procedure. The first measures will be taken for the analysis and data preservation for the internal use of the collaboration and for the open access to part of the data (Level 3).

After 1-3 years of the approval of the policy, the possible extensions to the Level 1 and Level 2 should be investigated. For Level 3 and 4, the reprocessing of the old data and the recording of the analysis workflows should become a routine procedure. The first public release of the CMS Level 3 data will take place in the long LHC machine shut-down of 2013-2014.

In long term, the adopted modus operandi should be monitored and evaluated. The planning for the long-term data preservation beyond the life-time of the experiment should start. For the Level 1 (additional data), the practice of providing and sharing additional data will have become a routine procedure. At the Level 2 (simplified data), the activities already in place and those implemented during the short-term period will be continued and consolidated. A persistent third-party storage for the data will be implemented. For the Level 3 (reconstructed data), the data preservation and access practices implemented during the first phase will be monitored and consolidated. The planning for the long-term data preservation beyond the life-time of the experiment starts. A regular sanity check for the preserved data at Level 4 (raw data) will be implemented. The open access, if endorsed by the Collaboration Board after the evaluation of the first public release, will become a routine procedure.

6. Benefits and frequently asked questions

The CMS concept of open access is based on the principle that the reconstructed Level 3 data, which will be made publicly available, is the same used internally in the CMS collaboration. This will guarantee that most of the effort for preparing the public data (reprocessing the raw data with the current version of the reconstruction software, documentation) will benefit the CMS collaboration. Making data public ensures that they will be available internally and open access acts as a driving force to make the data preservation to happen.

While open access to scientific data is a common practice in many branches, it is a new concept in HEP. The CMS experience with providing Level 2 simplified data samples for outreach and education has been very positive. The samples have been used in various occasions, even outside the field of particle physics, and the feedback has been enthusiastic. It can be expected that providing more open data can only be beneficial to the public image of the CMS collaboration.

Being a new concept in HEP, open access raises many concerns and they were taken into account in the final version of the policy. One of the main unknowns is the eventual additional workload to the collaboration. This may rise from a misinterpretation of the data in which case it was feared that the CMS collaboration would need to spend time and effort in verifying wrong claims based on the public data. It is true that by the time of the first public release of the early collision data, CMS will have a solid set of scientific publications on this data set covering the processes which can be studied and therefore leaving little space on any claims worth arguing, but only experience with public data releases will tell whether such situations arise. Another potential source of additional overhead is the support needed for the users of the public data. It is obvious that CMS will need to provide adequate documentation on the use of these data, but this effort will be useful also for the CMS collaborators. Furthermore, to protect the collaboration from external use cases that could generate burdens beyond the funding and resources, the policy states that the level of support that CMS will be able to provide to external users depends on the available funding.

Concerns were also expressed on the public availability of the analysis software. CMS will release, together with the data, the software tools to perform the analysis. It is however not intended to release the analysis-specific software, which will reside in an internal repository.

A question was also raised whether the public access will be competing with the internal computing resources of the collaboration. The CMS resources are estimated and funded according to the needs of the collaboration. Any important increase dedicated to open access will need to be funded separately and in addition.

The actual use and resources needed for open access are difficult to predict. For this purpose the current policy is a first stress-test exercise with a public release of a part of 2010 data in 2013 after which the experience is reviewed. In absence of unexpected overhead the public data release can be accepted as a standard procedure. If needed, the open access parameters can adjusted at that time.

7. Conclusions

The CMS collaboration has formulated a policy on data preservation, re-use and open access. The policy defines the principles with which the different levels of data are preserved for internal re-use and open access. The open, collaboration-wide discussion while preparing the policy was beneficial and concerns expressed during the preparation were taken into account in the final version of the policy.

CMS already provides some data for public use (numerical data in addition to the publications, simplified data for outreach and education) and it is planned that the first public release of reconstructed data will take place in 2013-2014. The CMS approach lies on the principle that the public reconstructed data is the same as the data used internally in CMS, therefore optimising the effort and resources needed for open access. The implementation of the policy is now starting, and the CMS collaboration is looking forward to see the interest and benefit raised by this initiative.

Acknowledgements

Authors wish to acknowledge the excellent assistance of Salvatore Mele and Sunje Dallmeier-Tiessen from CERN Open Access group when preparing the policy document, and the encouragement and support from the CMS management, especially from Joe Incandela, Teresa Rodrigo Anoro and Matthias Kasemann. Numerous colleagues have been of great help in discussing this topic and the feedback from the collaboration ensured that the policy now approved by the collaboration reflects the will of the CMS collaboration as a whole.

References

[1] DPHEP - ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics: https://www.dphep.org/

[2] CMS Collaboration, “The CMS experiment at the CERN LHC”, JINST **03** (2008) S08004

[3] CMS Collaboration, “CMS data preservation, re-use and open access policy”: https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=6032

[4] DPHEP-2009-001, “Data Preservation in High Energy Physics”: http://arxiv.org/ftp/arxiv/papers/0912/0912.0255.pdf

[5] CERN Document Server: http://cdsweb.cern.ch/

[6] CMS Collaboration, “Transverse momentum and pseudorapidity distributions of charged hadrons in pp collisions at sqrt(s) = 0.9 and 2.36 TeV”, JHEP **02** (2010) 41

[7] The Durham HepData Project: http://durpdg.dur.ac.uk

[8] INSPIRE Database: http://www.projecthepinspire.net/

[9] CMS Public Data: http://cms.web.cern.ch/org/cms-public-data

[10] S Malik, K Lassila-Perini, B Hegner, A Vedaee and M Stankevicius, “A Perspective of User Support for the CMS Experiment”, J. Phys.: Conf. Ser. **331** 082006

S Malik, K Lassila-Perini: “An outlook of the user support model to educate the users

community at the CMS Experiment", Proceedings of the DPF-2011 Conference, Providence, RI, August 8-13, 2011, http://arxiv.org/abs/1110.0355v1
[11] The CMS Offline Workbook: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBook
[12] The CMS Offline SW Guide: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuide
[13] K Lassila-Perini, S Malik, B Hegner, A Hinzmann and R Wolf, "Planning and organization of an e-learning training program on the analysis software in CMS" J. Phys.: Conf. Ser. **331** 082010
[14] CMS Collaboration, "CMS Computing TDR", CERN-LHCC-2005-023
[15] Worldwide LHC Computing Grid, Memorandum of Understanding: http://lcg.web.cern.ch/lcg/mou.htm