Version 1.2
Approved by the CMS Collaboration Board 20th April 2018

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

CMS data are unique and are the result of vast and long-term moral, human and financial investment by the international community. There is unique scientific opportunity in re-using these data, at different level of abstraction and at different points in time[1]. This opportunity calls for our collective responsibility, and poses unprecedented challenges as no data sample of this complexity and value has ever been preserved or made available for later re-use.

The CMS collaboration is committed to preserve its data, at different levels of complexity, and to allow their re-use by a wide community including: 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.

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 but relatively short embargo period, allowing CMS collaborators to fully exploit their scientific potential.

This policy describes the CMS principles of data preservation, re-use and open access, as well as the relevant actors in all these tasks and their roles and responsibilities. CMS understands that in order to fully exploit all these re-use opportunities, immediate and continued resources are needed. The level of support that CMS will be able to provide to external users depends on the available funding. This policy addresses the moral responsibility of CMS for its data, as well as the increasing concern of funding agencies worldwide and the civil society for the preservation and re-use of scientific data.

Notwithstanding the long-term perspective of the LHC programme, the time for action is now: lower-energy and lower-luminosity LHC runs at centre-of-mass energies of 0.9, 2.36, 2.76, 7 and 8 TeV may never be repeated, and their preservation and preparation for later re-use, has to be addressed urgently. Meeting this challenge is a unique way to stress-test and evaluate the entire preservation, re-use and open access concept for the CMS data.

CMS data take many forms. Starting from either raw experimental or simulated data through to reconstructed data and the datasets of higher abstraction generated by analysis workflows, and finally all the way to data represented in scientific publications. Each of these layers has the potential to afford different opportunities for long-term re-use and poses different challenges for preservation. Data represented in publications can already be preserved by building on the existing practices of the Collaboration (e.g. open access publishing) and existing third-party platforms (e.g. INSPIRE[2]), simply expanding the concept of publication to include additional data sets of a high level of abstraction. At the other extreme of the spectrum, closer to the raw data, different challenges appear which imply a paradigm shift from in-depth documenting and archiving of analyses during the publication process, to a preservation of reconstruction and simulation software packages with all of their dependencies.

[1] A. Holzner, P. Igo-Kemenes, S. Mele, "First results from the PARSE.Insight project: HEP survey on data preservation, re-use and (open) access" http://arxiv.org/abs/0906.0485

[2] http://inspirehep.net

In general, four levels of complexity of HEP data have been identified, which map on to re-use opportunities and preservation challenges[3]. These levels, adapted to the CMS context, allow one to frame the different approaches of this policy for the preservation and re-use of CMS data.

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).

CMS level 1 policy: CMS publishes its scientific results with open access journals. CMS strives to provide additional numerical information to facilitate immediate re-use and the combination of these results. This information is provided through, and archived in the long-term, by trusted third-parties such as INSPIRE and HEPData[4].

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.

CMS level 2 policy: CMS makes data of such high-level of abstraction available, accessible and preserved through CERN Open Data[5] and HEPData. INSPIRE counts their citations.

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.

CMS level 3 policy: CMS preserves the reconstructed data and simulations by keeping available a copy of the data reconstructed with the best available knowledge of the detector performance and conditions for each period of data-taking. A virtualised computing environment, compatible with the software version with which the original data can be analysed, is provided and maintained. In parallel, studies are conducted to develop simplified data formats common to all datasets and independent of specific software versions, adapted for physics analysis in the long-term future. Analysis procedures, workflows and code are preserved as part of the CMS code repository under the responsibility of the CMS Offline project. Responsibility for archiving of the data sits with the present tiered structure of the CMS computing infrastructure, while CERN Open Data serves as the publishing platform for selected data and is foreseen to take over preservation responsibilities in the mid to long term.

Level 4: Raw data and the software and documentation needed to access, reconstruct and analyse them

CMS level 4 policy: Any version of the CMS reconstruction and analysis software is required to be compatible with the raw data from the start of the data-taking. A custodial copy of the data is stored at CERN and at the corresponding custodial Tier 1 for that dataset. CMS software is released under an open source license and CMS will ensure that “custodial” copies of the software are also kept and are freely available.

[3] ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics, DPHEP, http://dphep.org. In particular C. Diaconu *et al.*, “Data Preservation in High-Energy Physics” http://arxiv.org/abs/0912.0255

[4] http://durpdg.dur.ac.uk

[5] http://opendata.cern.ch

Several potential re-use scenarios exists for the CMS data: by CMS collaborators, by experimental physicists who did not participate in the data taking, by theoretical physicists now or in the future, by scientists from cognate disciplines, educational and outreach initiatives, and by citizen scientists. The CMS constitution already allows researchers to become CMS affiliates and have access to its data and resources without being a full member of the CMS collaboration.

CMS will provide open access to its data at different points in time with appropriate delays, which will allow CMS collaborators to fully exploit the scientific potential of the data before open access is triggered.

- At level 1, the additional data is made available at the moment of the publication.
- At level 2, simplified data format samples are released promptly as determined 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. The portion of the data which CMS will normally make available is 50% after 3 years from data taking, rising to 100% within 10 years, but the Collaboration Board can, in exceptional circumstances, decide to release some particular data sets either earlier or later.
- At level 4, small samples of raw data potentially useful for studies in the machine learning domain and beyond can be released together with level 3 formats.

The first data release of 2010 data took place in 2014, as a stress-test exercise of the entire preservation, re-use and access chain. This release was followed by a full analysis of the procedure, which was endorsed by the Collaboration Board in 2015, and regular data releases, accompanied by appropriate simulated data, each approved by the Collaboration Board, are now taking place.

For the widest possible re-use of the data, while protecting the Collaboration's liability and reputation, data will be released under the emerging standard Creative Commons CC0 waiver[6]. 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. The release of data could create a community of users which may be nurtured through regular events organized by CMS, allowing further monitoring of the data re-use.

This data preservation, re-use and access activity implies responsibilities across several bodies within and beyond the collaboration. The Collaboration Board will approve, uphold and possibly amend this policy. The Data Preservation Coordinator, nominated by the Collaboration Board, will propose to the Spokesperson and to the Collaboration Board the release of data for wider access. The proposal will specify quality, quantity and location of the data to be released. The Coordinator will oversee the implementation phase, setting up the infrastructural components and the procedures implied by the four levels. The Coordinator will assure that the policy is followed, both on an ongoing basis at a lower level (such as the implementation of standards of documentations for analyses and the production and release of additional data sets) as well as for the events triggering large-scale data release at a higher level. The Computing Resource Board will evaluate and monitor the cost of upholding and implementing this policy, which will be discussed by the experiment management and the Collaboration Board.

---

[6] http://creativecommons.org/publicdomain/zero/1.0. The CC0 license is designed for data, and allows re-use by anyone, under the responsibility of these final users. This is analogous to CMS open access articles which are published under similar licenses designed for text documents, such as Creative Commons licenses such as CC-BY.
The emerging standard practice in disciplines where data reuse is common expects that third parties cite the original author of the data (e.g. through DOI, Digital Object Identifiers, available through INSPIRE).