PROCEEDINGS OF SCIENCE

Paving the Way for Open Data: 10 Years of Releases from the CMS Collaboration

Julie Hogan[a,*] **on behalf of the CMS collaboration**

[b] *Department of Physics and Engineering, Bethel University,*
3900 Bethel Drive, Saint Paul MN, USA

E-mail: j-hogan@bethel.edu

The Compact Muon Solenoid (CMS) Collaboration at CERN has been a pioneer in the open data movement within high-energy physics. For over a decade, the CMS Open Data project has provided public access to high-quality, research-level particle physics data. This initiative is primarily driven by the goals of data and analysis preservation, ensuring that the valuable data collected by the CMS experiment remains accessible and usable for future generations of scientists and the public. The 10-year history of the CMS Open Data program spans from initial releases of Run 1 data in 2014 to the latest significant release of data from 2016. The CMS Open Data model provides not only data but also the necessary tools and knowledge for analysis. This initiative has a growing impact for researchers outside CMS, and the most recent releases offer new opportunities for engaging with collider physics.

The European Physical Society Conference on High Energy Physics (EPS-HEP2025)
7-11 July 2025
Marseille, France

[*]Speaker

PoS(EPS-HEP2025)598

1. The Case for Open Data

The practice of making scientific data publicly available is a cornerstone of reproducible and transparent research. One of CERN's mission goals is to promote open access to science by the general public. Since the advent of the CERN Open Data Portal, the Compact Muon Solenoid (CMS) experiment has been a pioneer in releasing research-quality data.

The CMS Open Data program is guided by the principle that open data are only of value if they are in use. Publishing data in a public storage repository is only the first element of a functional open data release. Experiments must determine which data they will release, where data will be hosted, and how data will be discovered by users. But in order to transform this accessible data into reusable data, experiments must also provide metadata that document the contents of each dataset, software that is capable of processing the data, and instructions so that users can employ it confidently in their research. As noted by Matthew Strassler and Jesse Thaler in Nature [1], the steady publication of LHC data provides numerous benefits: it promotes the timely archiving of data and the associated knowledge before the collective memory of the collaboration fades; it empowers external scientists to analyze the data using new strategies; and it promotes early identification and correction of any deficiencies in the archival methods.

2. The CMS Open Data Model

To ensure the usability of its public data, CMS has developed a comprehensive open data model built on three pillars: Data, Tools, and Knowledge.

- Rich data in research-quality formats: CMS releases collision data, simulation, and occasionally "derived" datasets to the public. For all of these data types, CMS has taken the approach of releasing the same data format used by the majority of the collaboration for any given period of data collection. These formats are robust and would in principle allow replication of most CMS physics papers. Comprehensive metadata and supplementary information that enable proper analysis are made public alongside the data they describe.

- Robust software tools: high energy physics datasets have unique structures that aren't easy to interpret using typical industry-focused software environments. Many CMS data formats require a software framework called "CMS SoftWare" (CMSSW), and this is provided to users via either Virtual Machines or Docker Containers. This makes CMSSW available across operating systems and computer architectures. The lightest CMS data format, introduced during Run 2 of the LHC, can be analyzed directly with CERN's ROOT software or various python libraries, which are also provided in standalone Docker containers.

- Personable knowledge transfer: a multi-faceted pedagogical approach is critical to helping users break down barriers to unfamiliar data. Since 2020, CMS has offered regular workshops in which researchers can learn how CMS data are formatted, how to use either the CMSSW or python software containers to analyze the data, and basic guiding principles for data analysis. These workshops are coupled with the CMS Open Data Guide website, example analyses available on the Open Data Portal, and a help forum to connect users to CMS open data experts. Knowledge resources across modalities helps promote user success so that the CMS Open Data program can fulfill its mission.

This holistic approach ensures that researchers, students, and enthusiasts have everything they need to engage with the data meaningfully.

3. A Decade of CMS Open Data

The CMS Open Data program began in November 2014 and has continued with regular updates, progressively making more of the experiment's data public. The CMS Data Preservation policy [2] aims to make 50% of the data from a given data collection period available after six years and 100% after ten years. Figure 1 shows a summary of the CMS luminosity that is available on the Open Data Portal, and Fig. 2 documents the steady progression of releases over the past decade.

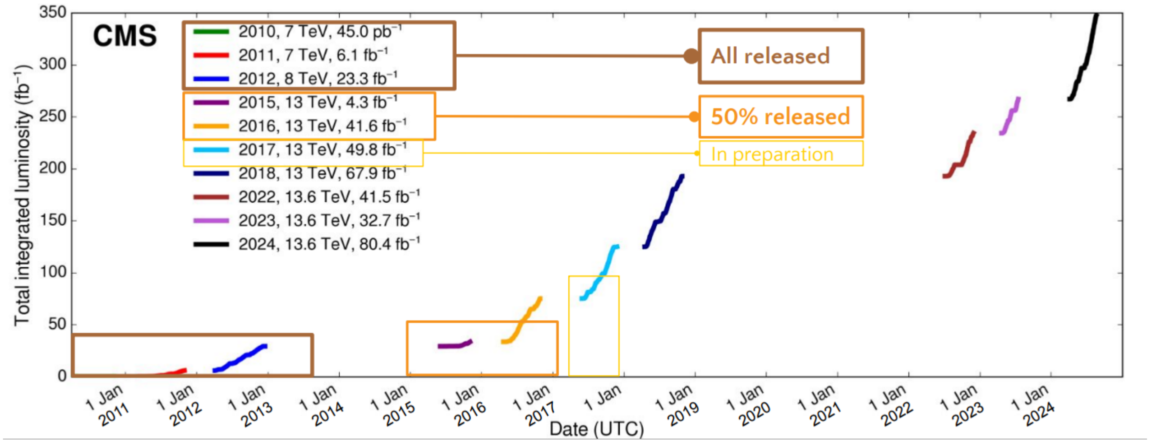

Figure 1: Cumulative luminosity versus day delivered to CMS during proton-proton collisions at the standard center-of-mass energy for each data collection period. This is shown for data-taking in 2010 (green), 2011 (red), 2012 (blue), 2015 (purple), 2016 (orange), 2017 (light blue), 2018 (navy blue), 2022 (brown), 2023 (light purple), and a portion of 2024 (black). The overlaid boxes indicate which proton-proton collision data have been released as open data. Figure adapted from [3].

The initial releases in 2014–2017 focused on data collected during Run 1 of the LHC. The earliest releases included 50% of the proton-proton (pp) collision data from 2010, 2011, and then 2012. The data were supported by a virtual machine environment to facilitate analysis, and after the first release simulations and examples were featured alongside the data. An iconic example to replicate the Higgs boson discovery plot was published [4], allowing users to investigate the histograms or dive into the CMSSW program to understand the event selection criteria. The data format in Run 1 is called the "Analysis Object Data" (AOD) format. The smaller luminosities of Run 1 allowed this analysis format to retain low-level detector information for each physics object, but required significant processing to apply certain calibrations and perform a physics analysis.

In 2019–2020, CMS released the remaining data from 2010 and 2011. This period brought more advanced tools for data analysis such as Docker containers, Jupyter notebooks, and additional analysis examples. CMS also began to release special samples dedicated to machine learning studies

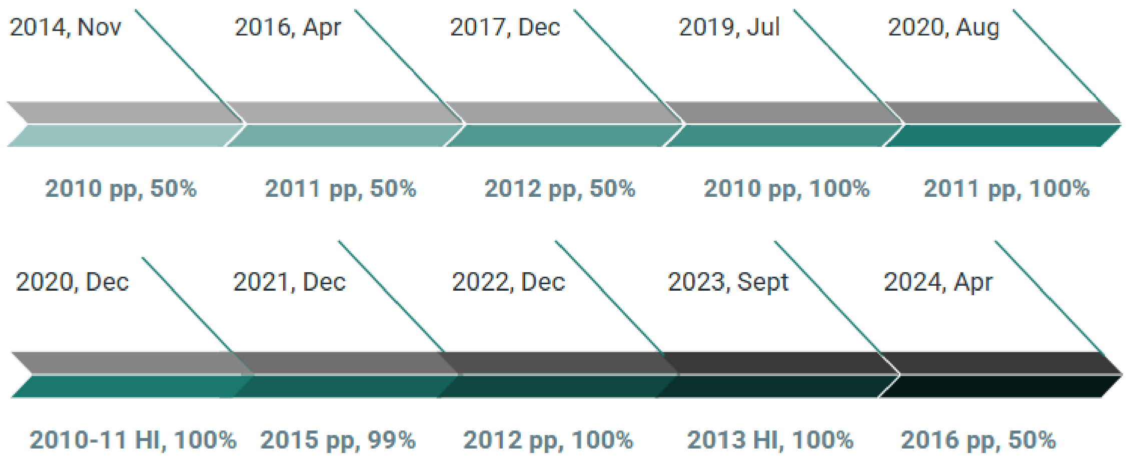

Figure 2: Timeline of CMS open data releases, summarizing the amount and type of data released.

and provided tools to create new simulations following the CMS workflow. The year 2020 saw the first CMS Open Data Workshop in October, and a complete release of heavy ion collision data from 2010–2011 in December.

The first data from Run 2 (2015 pp collisions) was released in December 2021 in a slimmer format called “MiniAOD" that simplifies analysis by applying many calibrations and algorithms to identify physics objects, removing the need to store a large amount of detector information. By the end of 2022, the full Run 1 pp dataset was made public, followed by the remaining Run 1 heavy ion data in 2023. A major advancement occurred in April 2024 with the release of a significant portion of the 2016 pp data (35.9 fb^{-1} total collected, 50% released) in the new, highly user-friendly “NanoAOD" format that does not require CMSSW for analysis. As of mid-2025, all data from 2010, 2011, 2012, and 2015 have been fully released. Data from 2017 are currently in preparation for future release.

4. Impact and Use Cases

The success of an open data project can be measured by its adoption and use within the scientific community. Since all CMS open data records are assigned unique Digital Object Identifiers, their usage can be tracked over time. As of September 2025, there has been a steady and near-linear increase in the number of scientific papers citing CMS open data, as seen in Fig. 3 (left), with the total number of papers surpassing 80. The vast majority of these papers are from small groups, demonstrating the project’s success in enabling research outside of large collaborations. Three examples from published research using CMS open data showcase its wide range of applications:

- Studies of quantum chromodynamics: Ref. [5] from 2017 was the first novel physics analysis performed using CMS open data. The authors studied the $2 \to 1$ splitting function of quantum chromodynamics by using jet substructure to decouple this function from other effects. They analyzed data from 2010 by processing CMS AOD files into a custom text file format.

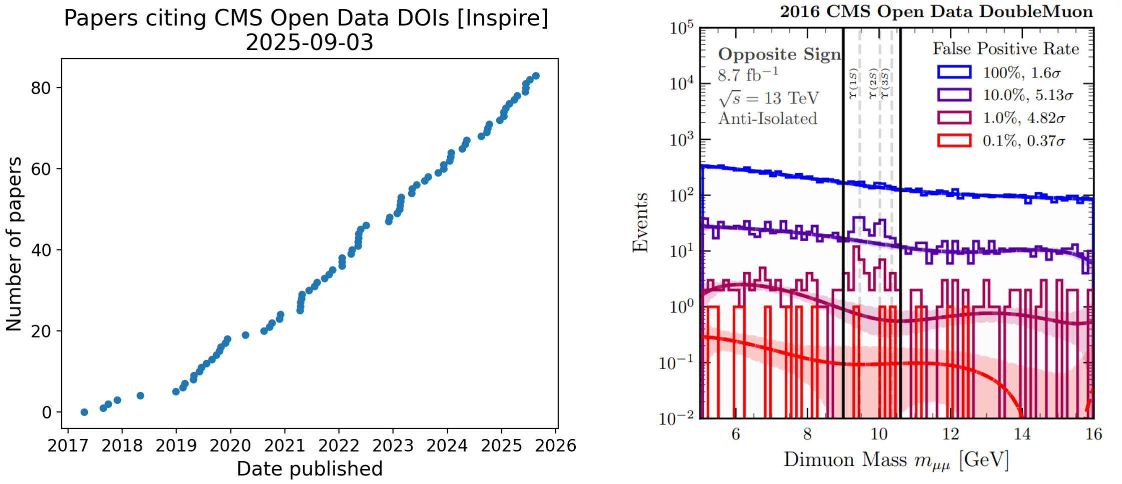

Figure 3: Publications citing CMS open data over time, collected from INSPIRE-HEP (left, figure from Ref. [8]). Distribution of muon pair mass showing various requirements on an anomaly-detection network, some of which allow the ϒ signal to be seen clearly (right, figure from Ref. [7]).

- New physics searches: Ref. [6] from 2021 is an example of a new physics search carried out using CMS open data. The authors searched for data matter in association with a heavy gauge boson called Z' using CMS data from 2012.
- Algorithm Development: Ref. [7] from 2025 utilized the latest release of 2016 data. This paper demonstrates how anomaly-finding neural networks can be used to uncover extremely difficult signatures. In this case, ϒ mesons are "re-discovered" using muon pairs that *fail* traditional muon isolation criteria. The anomaly detection technique is nevertheless able to reject sufficient background that the meson can be observed (Fig. 3, right). Many other articles have used CMS open data to test the development of new software programs or machine learning applications.

These examples underscore the value of providing research-grade data to the public, fostering innovation and enabling a broader scientific discourse.

5. New Opportunities and Future Outlook

The recent release of 2016 data opens up new avenues for research. The April 2024 release provided the first significant sample of 13 TeV proton collision data from 2016. This release includes over 16 fb^{-1} of collision data, a vast array of over 20,000 unique simulated processes, and updated software tools in the form of containers and virtual machines.

A key feature of the recent 2016 data release is the availability of the NanoAOD format [9]. This format represents a significant step toward making CMS data more accessible and reusable. The NanoAOD files are about 95% smaller than previous MiniAOD format (typically 1–2 kb per event). While earlier formats stored data using custom types that could only be interpreted by CMSSW, NanoAOD files store data in a more typical "tuple" format using standardized data types. This streamlined format allows data to be extracted directly into arrays using either ROOT or

PoS(EPS-HEP2025)598

Python, simplifying analysis workflow considerably. Their small size makes them more convenient to access over the internet, making it more feasible to avoid downloading any open data files locally.

To support its diverse user base, CMS provides a wealth of resources through the CERN Open Data Portal. For students and enthusiasts, resources include interactive event displays (including some in virtual reality!), materials for the International Masterclass program, and tools for university-level courses. For researchers, CMS offers comprehensive "Getting Started" guides for each data format, a separate webpage titled the “CMS Open Data Guide" that provides in-depth analysis instruction, and regular workshops where experts can connect with open data users and provide instruction and support.

The CMS Open Data project is a dynamic and ongoing effort. The collaboration has a clear roadmap for future releases, beginning with a partial release of the 2017 dataset that is planned for this year. By 2030, all data from Run 2 is planned for release, alongside the first releases of data from Run 3. The team will continue to update documentation, develop new analysis examples for Run 2 data (especially using NanoAOD), and expand outreach efforts using open data.

Over the past ten years, the CMS Open Data program has successfully established a robust framework for preserving and sharing high-energy physics data. By providing not just data but also the tools and knowledge required for analysis, CMS has empowered a global community of researchers and educators, amplifying the scientific impact of the LHC and setting a new standard for open science in the field.

References

[1] M. Strassler and J. Thaler, *Slow and steady*, *Nature Physics* **15** (2019) 725.

[2] CMS Collaboration, *2020 CMS data preservation, re-use and open access policy*, 2020. 10.7483/OPENDATA.CMS.1BNU.8V1W.

[3] CMS Collaboration, “CMS Luminosity – Public Results.” https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[4] N.Z. Jomhari, A. Geiser and A.A.B. Anuar, *Higgs-to-four-lepton analysis example using 2011-2012 data*, 2017. 10.7483/OPENDATA.CMS.JKB8.RR42.

[5] A. Larkoski, S. Marzani, J. Thaler, A. Tripathee and W. Xue, *Exposing the QCD Splitting Function with CMS Open Data*, *Phys. Rev. Lett.* **119** (2017) 132003 [1704.05066].

[6] S. Elgammal, M. Louka, A.Y. Ellithi and M.T. Hussein, *Search for the production of dark matter candidates in association with heavy dimuon resonance using the CMS open data for proton–proton collisions at* $\sqrt{s} = 8$ *TeV*, *Eur. Phys. J. Plus* **138** (2023) 548 [2109.11274].

[7] R. Gambhir, R. Mastandrea, B. Nachman and J. Thaler, *Isolating Unisolated Upsilons with Anomaly Detection in CMS Open Data*, *Phys. Rev. Lett.* **135** (2025) 021902 [2502.14036].

[8] T. McCauley, *cms-dpoa/data-usage: 03-09-2025*, 2025. 10.5281/zenodo.17046810.

[9] CMS collaboration, *NANOAOD: a new compact event data format in CMS*, *EPJ Web Conf.* **245** (2020) 06002.

PoS(EPS-HEP2025)598