Chapter 8

Early Experience with Open Data from CERN's Large Hadron Collider*

*Achintya Rao^{||}, Sünje Dallmeier-Tiessen[‡],
Kati Lassila-Perini[¶], Thomas McCauley[†]
and Tibor Šimko[§]*

[†]*University of Notre Dame, USA*

[‡]*Scientific Information Service, CERN, Switzerland*

[§]*Department of Information Technology, CERN, Switzerland*

[¶]*Helsinki Institute of Physics, Finland*

^{||}*University of the West of England, Bristol, UK
and CMS Experiment, CERN, Switzerland*

This chapter covers perspectives from the various partners who have worked on the release of large volumes of open research data from the Large Hadron Collider via the CERN Open Data Portal. The early experiences mentioned in the title refer to the launch of the Portal in November 2014 with the release of the first batch of high-level research data collected in 2010 by the CMS Collaboration. This chapter covers the motivation for releasing particle physics

*© 2019 The Author(s). Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY) 4.0 License.

data openly as well as the challenges faced in doing so and solutions developed to facilitate these efforts. The authors also touch upon the use cases of the open datasets and the impact the first release has had. Caveat lector: *The experiences and figures described in this chapter correspond to the year 2016 when this piece was originally written. The reader may want to consult Nature Physics 15 (2019) 113–119 to learn about later developments.*

1. Introduction

Located on the outskirts of Geneva on both sides of the Swiss-French border, CERN, the European Laboratory for Particle Physics, is the world's premier research facility for accelerator-based high-energy physics. The laboratory's flagship accelerator is the Large Hadron Collider (LHC), which, at 27 kilometres in circumference, is the biggest and most powerful particle accelerator ever built. It accelerates protons in clockwise and anti-clockwise directions to nearly the speed of light before colliding them at four points on its ring. Gigantic particle detectors, operated by international collaborations of scientists and engineers, are located at each of these collision points. They record information from these collisions, generating an enormous volume of data for analysis: each collision event produces data of the order of megabytes (MB), and, with the LHC delivering collision events 40 million times each second, the CERN Data Centre has collected tens of petabytes (PB) since the accelerator began operations in 2010.

The collisions produced by the LHC allow physicists to study the fundamental particles and forces of nature, and test the various theories and models that have been proposed to explain their behavior. Members of the LHC collaborations perform this research by analyzing the collision data and share the results in open-access publications. These datasets are truly unique in many ways, and, outside of the particle physics community, not only are they of interest to students and data scientists, they also form an important part of the scientific legacy of the LHC.

Now, a large portion of these data are being made publicly accessible without any restrictions to anyone in the world with an internet connection. In 2014, the CMS Collaboration released 27

terabytes (TB) of data that were recorded in the second half of 2010, representing half that year's data harvest.

Following this successful release, and equipped with the lessons learnt from the experience, CMS released a second batch of data in April 2016¹ (300 TB in total, including 100 TB of collision data recorded in 2011). However, this chapter focuses on the first release from November 2014.

2. Motivation

The release of open data from the LHC is motivated, in part, by the desire to ensure that these data continue to remain available to researchers in the future. The four LHC collaborations have therefore adopted policies for data preservation and access.² These policies include related matters such as embargo periods, licensing and reuse. The specifications in the data preservation policies distinguish four different data levels that help define the resulting recommendations³:

1. Level 1 data comprises data that is directly related to publications which provide documentation for the published results;
2. Level 2 data includes simplified data formats for analysis in outreach and training exercises;
3. Level 3 data comprises reconstructed data and simulations as well as the analysis-level software to allow a full scientific analysis;
4. Level 4 covers basic raw-level data (if not yet covered as Level 3 data) and their associated software and allows access to the full potential of the experimental data.

The open data being discussed in this chapter refer to Levels 2 and 3.

¹<http://cms.web.cern.ch/news/cms-releases-new-batch-research-data-lhc>.

²<http://opendata.cern.ch/collection/Data-Policies>.

³See <https://arxiv.org/abs/0912.0255>.

CERN is a strong proponent of openness in research, having worked with publishers and other laboratories to ensure that all particle physics results from the LHC collaborations are published as open-access papers.⁴ More recently this paradigm has been expanded to cover Open Science more comprehensively.⁵ CERN has helped build tools and digital libraries to foster Open Science practices beyond the particle physics community, such as the widely used Invenio digital library software. Invenio forms the basis for projects such as ZENODO, which has helped thousands of scientists archive their work on CERN's servers with persistent identifiers. In order to facilitate storage for and access to open data from the LHC collaborations, the laboratory launched the CERN Open Data Portal, built using Invenio, in November 2014.

All four LHC experiments provide open data in a format suitable for a classroom environment, typically used in the Physics Masterclasses⁶ for high school students around the globe (see also Section 5.1). In addition to these “educational” datasets, the CMS Collaboration, in accordance with its data preservation and data access policies, has also released large volumes of high-quality data for research purposes along with the tools necessary for accessing and analysing them. This chapter describes the motivation behind the launch of the CERN Open Data Portal as well as the background to and experience with the first release of open data from the LHC.

3. Challenges

3.1. Volume of data

Making data public does not make them any simpler. Since the LHC began delivering collisions in 2010, and as of writing this document (June 2016), the four detectors have recorded several tens of PB of collision data. Performing an analysis on this volume of data involves

⁴See <https://scoap3.org/> and <https://cds.cern.ch/record/1955574>.

⁵<http://home.cern/cern-people/opinion/2014/11/road-open-science>.

⁶<http://physicsmasterclasses.org/>.

several stages. First, the data from the collisions, which are in the form of electrical signals from individual channels within each detector, must be pieced together, or reconstructed, to form a coherent picture of the collision itself. Then, the analysis software must sift through the collisions to look for signatures associated with the predicted physics phenomena or for deviations from the predictions.

The LHC collaborations rely on the Worldwide LHC Computing Grid (or the Grid, for short)⁷ to perform these analysis tasks. The Grid is a network of interconnected computers distributed all over the globe, and datasets are broken into chunks to facilitate processing. While those outside the LHC collaborations do not typically have access to these resources, it is worth pointing out that the research-level open data from the detectors are already “reconstructed” and “reprocessed” (Level 3 data) — that is, a lot of the gruntwork has already been done to convert the datasets into a form useful for the final analyses themselves (see also Section 6.2). Despite this, performing research with the “analysable” data is no small task and preparing an appropriate dataset, depending on the analysis one wishes to perform, can take up considerable time and computing resources. To make life easier, the Portal provides smaller, simplified datasets — known as “derived datasets” — that can be used for analysis in a classroom or university environment without a large technological overhead. The Portal also makes available the code used to prepare these derived datasets.

3.2. Software environment and documentation

Even in a format that includes meaningful objects reconstructed from raw data, the data from the LHC still remain very complex. The data files cannot be opened and understood as simple data tables — they contain special structures with all the necessary information for data interpretation — and dedicated software is necessary to analyse them. The software and an appropriate working environment must be provided together with the open data. Using these data for scientific purposes requires good knowledge of experimental

⁷<http://wlcg.web.cern.ch/>.

particle physics and functioning of a detector, and documentation can hardly cover all these aspects, understanding of which is usually acquired during undergraduate and graduate studies.

It is also important to remember that the release of the data happens some years after these data were collected. Roles within the collaboration change regularly and those responsible for various data-collection efforts at the time of data-taking are not necessarily involved in those activities come data release. Therefore, significant efforts are needed to acquire and capture the relevant documentation and preserve the institutional knowledge of the collaboration.

4. Solution

A collaborative approach was taken to deliver both the CERN Open Data Portal and the first big open data release — from CMS — at the same time. Members of CERN IT, CERN Scientific Information Service and the CMS Collaboration developed the platform, catalogued the data and analysis tools for the release, and prepared the appropriate user guidance and documentation. From the very beginning, it proved to be important to have subject specialists, information specialists and IT experts on board together to tackle the delivery of the big, complex datasets. Regular working meetings, coding/documentation sprints and feedback cycles ensured that the work on the Portal and its content was on track to meet the requirements of the potential users and use cases.

4.1. *Developing the platform*

The CERN Open Data platform is based on the Invenio digital library framework,⁸ which was originally developed at CERN and is now co-developed by an international collaboration.

Invenio provides an ecosystem of standalone independent packages that can be used to build custom digital repositories. Use cases for Invenio include integrated library systems, digital document

⁸<http://inveniosoftware.org/>.

repositories, multimedia archives and data repositories. The software framework has been used for numerous such instances around the world and is particularly known for its scalability for large libraries, such as CERN Document Server (CDS)⁹ or INSPIRE.¹⁰

Invenio is mainly used in academic environments. In order to make the CERN Open Data Portal attractive to the general public, including high school students and other targeted audiences, several features such as documentation pages and glossaries have been included on the Portal. Figure 1 shows the home page of the CERN Open Data Portal.

A crucial aspect of providing open data is to ensure the long-term preservation of the datasets, as well as the accompanying software tools. Invenio enables the persistence and longevity of the Portal's assets, drawing inspiration from recommended practices of the Open Archival Information System (OAIS).¹¹ As part of this,

Fig. 1. CERN Open Data Portal home page

⁹<https://cds.cern.ch/>.

¹⁰<https://inspirehep.net/>.

¹¹https://en.wikipedia.org/wiki/Open_Archival_Information_System.

persistent identifiers — in this case Digital Object Identifiers (DOIs) — ensure that the datasets and software tools can be referenced and cited well into the foreseeable future.

The datasets available on the CERN Open Data Portal are themselves stored in CERN's EOS¹² disk storage system. EOS is a disk-based service that provides low-latency storage infrastructure for physics use cases. EOS manages over 120 PB of disk storage at the time of writing this chapter. Access to the open data is provided by the XRootD¹³ protocol and is well suited to situations with many concurrent users, a significant fraction of random data access, and a large rate of file opening. The XRootD protocol allows *ad-hoc* streaming of necessary portions of data during an analysis, saving potential users of the large datasets from having to preemptively download the complete datasets before starting an analysis.

In the spirit of openness, development work for the Portal is conducted publicly on the code-sharing platform, GitHub.¹⁴

4.2. Cataloguing data and analysis tools

The CERN Open Data Portal has been developed to serve two main use cases: education and research (see more under Section 5). This is reflected in the information architecture of the Portal itself. It should be noted that special emphasis has been given to providing contextual information that is needed to understand and reuse the data and accompanying software.

On the Portal, each dataset has its own record. Related records are aggregated in dedicated collections, such as those for “Primary datasets” or “Derived datasets”. This aggregation, which is standard practice in digital libraries, has been adopted to research data.

¹²<https://eos.web.cern.ch/>.

¹³<http://xrootd.org/>.

¹⁴<https://github.com/cernopendata/pendata.cern.ch>.

Table 1. The MARC metadata model extended to describe the CMS primary datasets. Not exhaustive, showing some fields only

MARC		
Tag	Description	Example Value
001	Record ID	1
024	DOI	10.7483/OPENDATA.CMS.A342.9982
110	Corporate author	CMS Collaboration
245	Dataset title	/BTau/Run-2010B-Apr21ReReco-v1/AOD
250	Release	CMSSW_4_2_1_patch1
256	Dataset description	2916 files, 25423849 events, 2678962237287 bytes
556	Validation information	(link to list of validated runs)
567	Selection information	(description of how events were selected)
581	Usage instructions	(link to how to guide)
583	Certification	(description of certification and links to publications)
856	File information	(links to ROOT files stored in EOS)
980	Collection	CMS-Primary-Datasets

Further, metadata — standardised according to MARC¹⁵ — are provided within each record. Adopting the practices developed by partner information platforms in high-energy physics (e.g., INSPIRE and CDS), standard metadata fields were populated. However, various additional metadata fields needed to be created to make sure information about data are standardised and comprehensive across all the datasets available via the Portal (see Table 1).

The information provided was vetted to be comprehensive enough for future reuse of the data. Certain fields, such as the names of the authors, the publication year, title of the dataset, name of the publisher, were made mandatory for the sake of minting DOIs.¹⁶

¹⁵https://en.wikipedia.org/wiki/MARC_standards.

¹⁶See <http://doi.org/10.5438/0008>.

The metadata are currently stored as XML files, although it is planned to move to JSON schemas with the next release of Invenio (version 3). JSON schemas will provide more flexibility and interoperability with other platforms.

The data and metadata on the Portal are provided under a Creative Commons Public Domain Dedication (CC0). This liberal choice should enable any kind of reuse and exploitation of these datasets. Similarly, the software releases available via the Portal are also published under Open Source licences. The respective licences are displayed on the individual records on the Portal.

Two example records for primary data are shown in Fig. 2 for basic metadata, alongside the Datacite Metadata schema and Fig. 3 for more specific information such as validation procedures or limitations of the content displayed.

4.3. Preparing user guidance and documentation

As mentioned earlier, performing a physics analyses with the open data from the LHC is far from trivial. To lower the barrier of entry

Fig. 2. Primary data record on the CERN Open Data Portal

Fig. 3. Detailed metadata is provided to enable users to reuse the data easily

and make the datasets meaningful for students, citizen scientists and members of the wider particle physics community, clear and thorough documentation has been provided, either on the Portal itself or via external resources stored on CERN webpages.

Performing a high-level analysis on the datasets requires the use of a special software framework, known as ROOT,¹⁷ which was developed at CERN. In addition to ROOT, detector-specific software packages are needed to analyze data from the different LHC collaborations. For example, CMS uses a framework called CMSSW.¹⁸ To streamline access to these tools, the software environments necessary to perform a physics analysis on the LHC datasets are captured using the CernVM¹⁹ technology. The CernVM Virtual Software Appliance provides a complete, portable and easy-to-configure user environment for developing and running LHC data analysis locally or on institutional and commercial clouds (such as

¹⁷<https://root.cern.ch/about-root>.

¹⁸<https://cms-sw.github.io/>.

¹⁹<http://cernvm.cern.ch/>.

OpenStack, Amazon EC2, Google Compute Engine). Use of CernVM removes the need for separately installing the various software packages and libraries. Documentation on the Portal also guides the users through the basic actions needed to familiarise themselves with the tools provided.

Once the datasets, the software environments and the documentation have been assembled, an important challenge was to present this information in an easy-to-use way. A considerable effort was therefore devoted to an attractive presentation of assembled data and tools needed to work with them.

Further, precise instructions were provided for getting started with an analysis using the CMS data, by documenting the installation and testing of the applicable virtual machines and the initialisation of the CMS analysis environment. Prior to the release of the data, these steps were extensively tested on a variety of hardware and software configurations. A validation report, based on this testing, was also produced and made available on the Portal.

Yet another barrier, one that might not seem obvious to those used to working with particle physics datasets, is one of vocabulary. This is not limited just to the terminologies within particle physics; several terms are used in CMS-specific contexts, referring to the analysis software or to the information recorded in the datasets themselves. A Portal-wide glossary²⁰ was therefore created; terms included in the glossary are highlighted on all pages on the Portal, and hovering on the term displays the definition. This helps provide contextual information within the documentation.

5. Use Cases

The division of the CERN Open Data Portal into two top-level sections — Education and Research — reflects the two main anticipated use cases. The division also reflects the respective levels of physics and software knowledge needed to use the content in each

²⁰<http://opendata.cern.ch/glossary>.

section. Primary datasets from the LHC detectors found in Section 5.2 reflect the complexity of the detectors themselves. To be able to extract something meaningful from these datasets one needs at least: knowledge of the physics involved, familiarity with the data format, knowledge of the detector and its performance, a software environment and API specific to the experiment, and software expertise. These are significant hurdles for the general public. The material found in Section 5.1, on the other hand, has already found use by the general public in many cases and is more easily accessible.

5.1. Education

The first and most enduring use of public data from the LHC is in education, particularly in programmes designed to teach particle physics to students of high school age (in so-called Physics Masterclasses described in further detail below). Before open data policies were in place, requests for data to use in these programmes were done on an ad-hoc basis. The success of the educational programmes was one of the favorable factors considered by the LHC collaborations when further data releases and open-access policies were considered.

The main content under the Portal's Education section consists of derived datasets (mentioned in Section 3.1) and two browser-based applications for exploration of the derived datasets. Derived datasets are reductions of the primary datasets in the sense that only part of the information in the primary datasets is kept and/or only some of the collision events in the primary datasets are selected. Very often the derived datasets are produced in open, human-readable formats such as CSV, XML and JSON; this is possible when only some of the information from the primary datasets is kept. Software for production of derived datasets is included in the Portal, as are code examples for visualisation and analysis of the datasets.

The actual content and level of complexity of a particular derived dataset often depends on the intended audience. Open data as distributed in derived datasets find widest use in the aforementioned Masterclasses. Here, the intended audience are high school students.

The purpose of the masterclass is to teach students a bit about particle physics and the detectors used to study it, as well as to give them a sense of what it is like to analyse data and to obtain a result. In 2016, the International Particle Physics Outreach Group (IPPOG) Masterclasses coordinated by QuarkNet and TU Dresden were held in 46 countries with over 10,000 students participating between February and April.

A dataset used in a masterclass is prepared by a physicist with a specific analysis in mind. For example, both ATLAS and CMS have exercises where the students are meant to estimate the ratio of positively charge W bosons to that of negatively charged W bosons. Therefore, a sample of W bosons must be extracted from the primary datasets. These datasets are distributed as either CSV or XML files. A typical masterclass day starts with a short lecture on particle physics, how the detector works and how to conduct the analysis to follow. The students view up to 100 events in an interactive collision visualiser (called an event display): an application that visually renders the collisions and the detector. With the display, the students attempt to determine whether or not the event displayed is a W^+ or a W^- .

Direct exploration of the data is also possible on the Portal itself, via an interactive browser-based event display and histogram tool. The event display visualises derived datasets from CMS; a screenshot can be seen in Fig. 4. The histogram application allows the user to examine the distributions of variables found in several of the derived datasets. An example is shown in Fig. 5.

5.2. Research

Open access to the research data will, in the long term, allow the maximum exploitation of their scientific potential. The potential users of the Research collection are experimental and theoretical physicists who were not members of the collaboration collecting the data as well as collaboration members who might want to return to the data long after they were recorded. Also, these data can be of use

Fig. 4. The browser-based event display found in the CERN Open Data Portal. An event derived from the DoubleMu primary dataset is shown

Fig. 5. A screenshot of two histograms created using the application found in the CERN Open Data Portal

in other domains: interest in CMS open data has been expressed from areas such as machine learning and development of statistical methods.

To facilitate the data use, technical details regarding the information contained in the data, as well as key points to keep in mind when performing an analysis, were collected on the Portal. This includes explanations of so-called physics objects,²¹ which represent individual particles or groups of particles produced in an individual collision.

Reproducibility is a crucial aspect of the scientific method, and it cannot be achieved without open data. The reproducibility of CMS results using the provided open data has been tested with high-level validation analyses performed on the primary datasets available on the Portal. These validation benchmarks reproduce highly cited analyses that CMS published with a similar data volume.²² Although these analyses are not too complicated, they nevertheless have interesting physics contents, and offer the possibility of comparison with data at other beam energies in the future. All steps in these benchmarks are doable with the tools and information available publicly; information that was previously lacking to complete the process is being made available on the Portal.

Novel research is also being conducted using the CMS open data. Theoretical physicists at MIT recently presented new particle physics analyses that used CMS open data.

6. Impact

6.1. *Impact on general public*

The launch of the CERN Open Data Portal in November 2014 along with the publication of the CMS primary datasets (27 TB), derived datasets, interactive event visualiser, the CMS virtual machines and

²¹<http://opendata.cern.ch/about/CMS-Physics-Objects>.

²²See the validation record <http://opendata.cern.ch/record/460> and the corresponding CMS publication <http://dx.doi.org/10.1088/1748-0221/7/10/P10002>.

the analysis software examples was met with considerable interest from the general public.

The launch was covered by major news networks and on social media, with many articles inviting readers to run an LHC physics analysis “in your living room”. In the month following the release, more than 80,000 people visited the Portal, which is several times larger than the active particle physics community (estimated at around 20,000). Most of the visitors spent time browsing around the site and using the provided interactive tools. Several thousands of visitors explored the provided data by downloading the virtual machines and software to study the datasets.

Two weeks after the launch of the Portal, an “Ask Me Anything” session was organised on Reddit,²³ where readers could ask questions about open data and open science practices at CERN. The Reddit AMA session attracted considerable attention and led to another peak in visitors to the CERN Open Data Portal.

The public interest in the released open data generated some non-physics use case scenarios, such as using the large datasets as a robust input for digital forensics studies related to security in cloud computing. It is enlightening to see possible applications outside of the primary focus of the data, which is physics analysis.

6.2. Impact on scientists

When the CMS Collaboration first drafted its open data policy in 2012, the concept of open research data was entirely new in high-energy physics. The researchers — correctly — see the published result as the goal of the research work, and publishing data on their own was seen by some as unnecessary, premature or meaningless. Several concerns were expressed within the collaboration, mainly addressing the potential additional workload for collaboration members in case someone presents contradictory or false results based on the open data from CMS. It was, however, considered that the benefits of open data, both in terms of preserving their

²³<https://redd.it/2nxwkb>.

scientific potential and in enabling their use in education, overcome any potential risks.

The high-level open data from CMS — in the “primary datasets” — are in the same format as the ones used within the collaboration for physics analysis. There is no simplification or curation of these data for public release. However, in order to prepare these datasets in the first place, the raw data from the detectors must be “reprocessed” and the same version of the software used in the reprocessing must then be used to perform an analysis on the datasets. The only viable solution for releasing the reprocessed primary datasets with the appropriate software versions was to make them available as part of the CMS Collaboration’s legacy data. The legacy data for the 2010 data collection period are reprocessed in their entirety with a single version of the analysis software. In this way, there is no need for additional reprocessing and validation just for a public release, a time- and resource-consuming process. The exact provenance details (such as configuration parameters) can be directly extracted from the CMS internal data description systems.

With that being said, finding the human resources to provide adequate instructions and documentation is quite demanding for a collaboration during active data taking, with all the resources focused on collecting and analysing new data. The experience of the first data release from CMS was instructive: even though preparations for the data to be released were started with a fairly short delay after the final reprocessing of the legacy data, and considering that some of the last publications with these data were done just a bit earlier, finding the correct set of instructions and documentation was not easy for many simple reasons. For one, the people in charge had moved to different positions within the collaboration. For another, the documentation had been updated and, although much of it was version-controlled, finding the correct set of instructions corresponding to the exact software release is not as obvious as it may seem. Having learnt from this, CMS is preparing for smoother data releases in the future by recording all the necessary instructions for data analysis with the data that are in active use.

Beginning preparations for releasing open data well in advance has had a major impact on the long-term data- and knowledge-preservation efforts within the CMS Collaboration. Curating the instructions on how to use the data puts the emphasis on the most difficult area in data preservation: the preservation of knowledge, which at the time of the active use of the data is such an integral part of everyday work that it may not even have been explicitly noted down.

The overwhelmingly positive feedback received following the first data release in 2014 has convinced the CMS Collaboration of undertaking further regular data releases.