Preparations for the CMS 2016 Data Release

CMS Data Preservation and Open Access

CERN Summer Student Program - Work Project Report

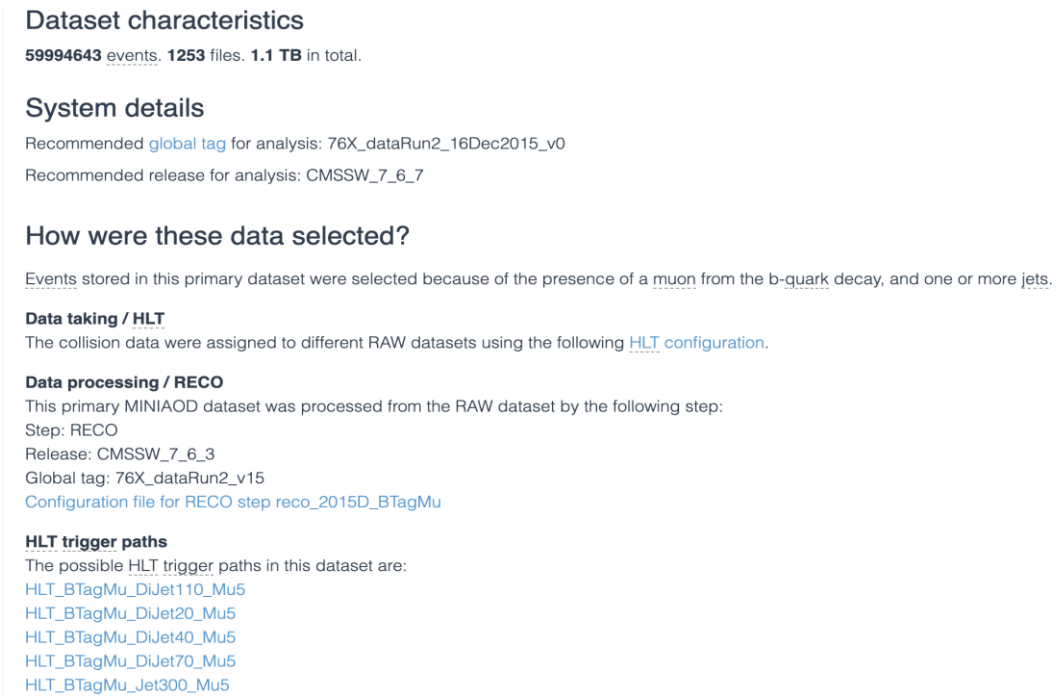
Nancy Hamdan - Joud Masoud

Supervisor: Kati Lassila-Perini

August 2023

Introduction

The CERN Open Data portal [1] serves as a hub for accessing a wide array of data generated from different experiments at CERN. The portal enables research-level use of the data by providing data records, their corresponding environment, software, and supplementary materials like the documentation necessary to comprehend and analyze the data. The records on the portal can be searched by different options including experiment, data type, and collision energy type. Moreover, aligned with internationally recognized norms in open science and data preservation, the portal distributes its materials using open licenses. Each material is assigned a Digital Object Identifier (DOI), enabling them to be cited as distinct and identifiable resources. This approach ensures transparency and accessibility in sharing the results of CERN's research efforts.



Dataset characteristics

59994643 events. **1253** files. **1.1 TB** in total.

System details

Recommended global tag for analysis: 76X_dataRun2_16Dec2015_v0

Recommended release for analysis: CMSSW_7_6_7

How were these data selected?

Events stored in this primary dataset were selected because of the presence of a muon from the b-quark decay, and one or more jets.

Data taking / HLT
The collision data were assigned to different RAW datasets using the following HLT configuration.

Data processing / RECO
This primary MINIAOD dataset was processed from the RAW dataset by the following step:
Step: RECO
Release: CMSSW_7_6_3
Global tag: 76X_dataRun2_v15
Configuration file for RECO step reco_2015D_BTagMu

HLT trigger paths
The possible HLT trigger paths in this dataset are:
HLT_BTagMu_DiJet110_Mu5
HLT_BTagMu_DiJet20_Mu5
HLT_BTagMu_DiJet40_Mu5
HLT_BTagMu_DiJet70_Mu5
HLT_BTagMu_Jet300_Mu5

Fig 1: Example of a dataset record on the CERN Open Data Portal

The CMS open data records are released on the CERN Open Data Portal and are accompanied with information such as content metadata describing a dataset's size, number of events, and number of files, contextual metadata describing what environment and software to use to analyze the data at a research level, and provenance metadata describing how the data was collected and reprocessed. Every CMS data release has its own set of data-curation scripts that are used to create the records which are released on the portal [2]. The scripts access different CMS services such as CMS DAS (Data Aggregation System) [3] and CMS McM (Monte Carlo request management system) [4] and collect information about the data-taking configuration files, HLT path listings, and the reprocessing information associated with the dataset. The scripts then extract and consolidate the necessary metadata information into the data records that are released on the portal. The data records can be downloaded using XRootD [5], direct HTTP download, or the cernopendata-client [6] command line tool which allows data inspection too.

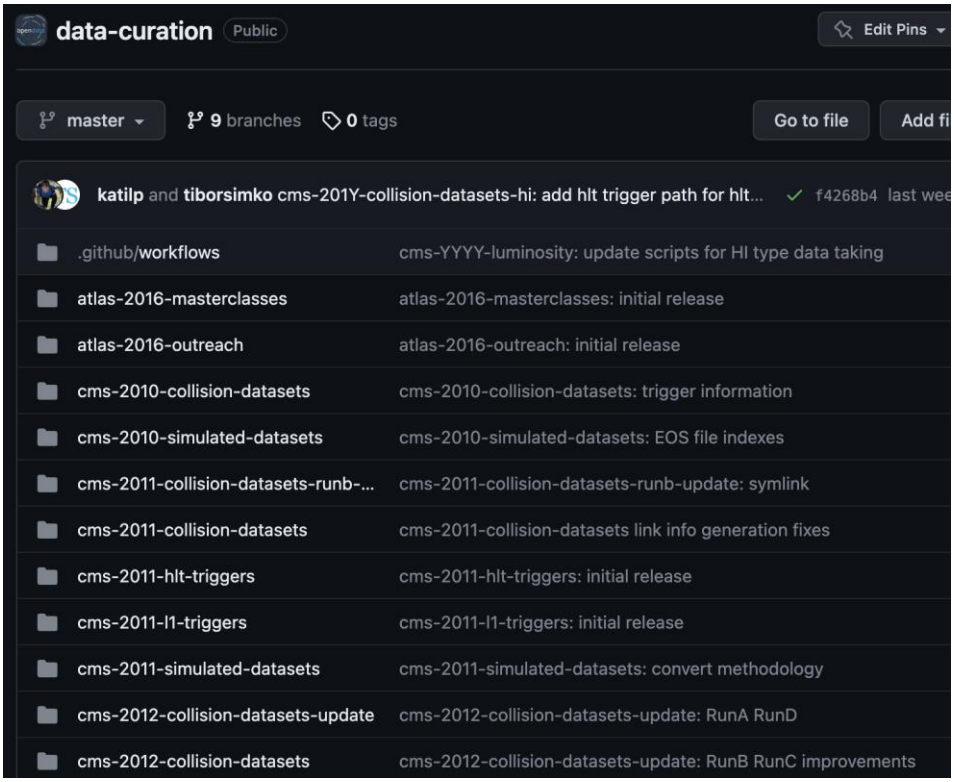


Fig 2: data-curation scripts as seen on GitHub and divided by data release

Project Details

Over the course of this summer, our project was centered around the preparations for the CMS 2016 open data release. This release includes both the 2016 collision datasets and the corresponding simulated datasets. Our efforts were concentrated on the preparation and development of the data-curation scripts essential to this release.

As part of our work on preparing the data-curation scripts for the 2016 release, we have worked on streamlining the metadata extraction processes in the scripts. This included:

- Improving the configurability of the scripts by better handling hardcoded values.
- Automating obtaining some of the metadata fields values like the date reprocessed and run numbers from DAS using the dasgoclient.
- Automating getting run period, global tag for processing, and the right configuration files that were used for processing.
- Automating extracting HLT listings for a specific year from the Run 2 HLT listings using a python script that could be used in the data-curation scripts of the remaining Run 2 data releases.

We have also worked on adding a new metadata field, ‘container_images’ which is a metadata field that can be used to get the fully qualified image name (FQIN) of the container image that is recommended for analyzing the data. The ‘container_images’ field is an array of possible instances of this recommended image that are hosted on different registries like Docker [7] or Gitlab [8]. Without this new metadata field, analysis workflows using CMS data had to use hardcoded values of the right container images FQIN that must be used which had to be known by the authors of the workflows.

To add ‘container_images’ as a metadata field, we updated the JSON schema that defines the structure and data type of the data fields that the records on the CERN Open Data Portal can have to include ‘container_images’.

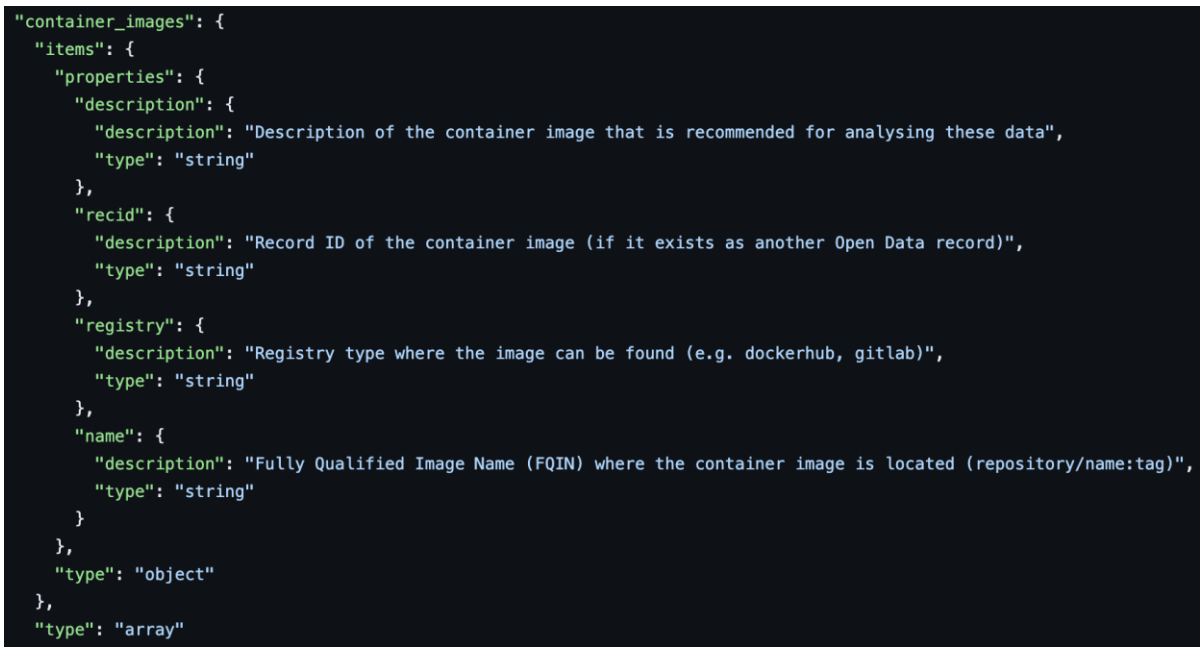


Fig 3: JSON schema of the ‘container_images’ metadata field

We also created a python script that updates existing records to include ‘container_images’. The script can also be used to automatically update the values of ‘container_images’ in existing records in case for example, a new registry has been added for where the container images can be found at, or changes were made to the fully qualified image names of the images on some registry.

In the process of adding the new metadata field, adjustments were made to the cernopendata-client to accommodate these updates. This involved introducing an additional option called "--filter" to the "get-metadata" command. This enhancement was designed to align with user requirements, simplifying the process of extracting metadata information using the client. With this new option, users can retrieve a single value from a metadata array. Prior to the addition of this option, extracting information stored in arrays was not achievable using the client.

```
$ cernopendata-client get-metadata --recid 329 --output-value authors.name
Adam-Bourdarios, Claire
Cowan, Glen
Germain, Cecile
Guyon, Isabelle
Kégl, Balázs
Rousseau, David
$ cernopendata-client get-metadata --recid 329 --output-value authors --filter orcid='0000-0002-9764-9783'
{
    "affiliation": "INRIA, Saclay; LRI, Paris 11; Orsay",
    "name": "Germain, Cecile",
    "orcid": "0000-0002-9764-9783",
    "rorid": "02kvxyf05; 04e3ktk27; 028rypz17"
}
$ cernopendata-client get-metadata --recid 329 --output-value authors.name --filter orcid='0000-0002-9764-9783'
Germain, Cecile
```

Fig 4: Example on using the ‘--filter’ option in cernopendata-client

We had to make sure that ‘container_images’ is displayed correctly on the open data portal as part of the records’ details. To do this, the html template which is used to display the pages of the records on the open data portal was updated to display the ‘container_images’ field.

As a last step of adding ‘container_images’ as a new metadata field, we documented all of the steps involved in adding a new metadata field ranging from modifying the JSON schema to updating the cernopendata-client and displaying the new metadata field on the portal. We added this documentation to the CMS open data release guide.

Plans for the Future

During our project this summer we have finished preparing the data-curation scripts for the CMS 2016 collision datasets. In the future and as part of our continued work on the CMS 2016 data release, we will prepare the data-curation scripts for the 2016 simulated datasets while implementing the lessons we learned from preparing the data-curation scripts of the 2016 collision datasets, and we will further enhance the data-curation scripts to improve their reusability.

Changes in CMS systems like DAS and McM that happen over the years make it hard to reuse the same data-curation scripts of one release and their methods of retrieving information from CMS systems for another release. Because of this, it is important to ensure that the data-curation scripts are as reusable as possible so that future releases can happen quicker and more smoothly. It is also important to ensure that the CMS open data release process including using the data-curation scripts is well documented such that it is easy to onboard new people to work on future releases. In light of this, our long-term goals for our work on this project include:

- Improving automation: better utilize the command line tools of the different CMS services to get metadata values such that the number of remaining hardcoded values is minimal.
- Improving configurability: separate hardcoded values and configuration values that are release specific from the remaining code such that they are easier to change for future releases.
- Improving code modularity: modularize as much logic of the scripts as possible such that more parts of the scripts can be easily reused in the future.
- Improving error handling and logging: add more checks and error logging in parts of the scripts that query CMS services such that if the services change in the future the scripts gracefully handle failures and report them well.
- Improving code documentation: add more comments to the code that better explain how the different parts in it are working.
- Utilizing the cernopendata-client better: separate actual metadata from metadata that is used for display purposes only and use the client to get the metadata that is for display purposes.
- Continue and improve CMS release guide: complete the documentation on the process of releasing simulated data, improve existing documentation on the release of collision data, and add more examples across the guide on how to use the data-curation scripts.

Acknowledgements

We are grateful for the invaluable guidance and support provided by our supervisor, Kati Lassila-Perini, whose expertise, and mentorship have been instrumental throughout our internship journey. We extend our appreciation to our university, Princess Sumaya University for Technology, for their encouragement and the opportunity to partake in this enriching experience. Additionally, we would like to express our gratitude to the CERN and Society Foundation for their contributions that made this internship possible, and to the CERN Summer Student Program for fostering an exceptional learning environment.

References

[1] CERN Open Data Portal, https://opendata.cern.ch/ (2021), accessed: 2023-08-23

[2] Cern Open Data, Data Curation Scripts, https://github.com/cernopendata/data-curation/blob/master/README.rst, accessed: 2023-08-17

[3] Kuznetsov, V., Metson, S., & Evans, D. (2010). The CMS data aggregation system (No. CMS-CR-2010-036). https://doi.org/10.1016/j.procs.2010.04.172

[4] G. Boudoul et al, “Monte Carlo Production Management at CMS” J. Phys.: Conf. Ser. 664 (2015)

[5] XRootD software framework, https://xrootd.slac.stanford.edu/ (2020), accessed: 2023-08-17

[6] CERN Open Data Client, https://cernopendata-client.readthedocs.io/en/latest/ (2023), accessed: 2023-08-22

[7] Docker, https://www.docker.com/ (2023), accessed: 2023-08-17

[8] Gitlab, https://about.gitlab.com/ (2023), accessed: 2023-08-17