## **THE CERN ANALYSIS PRESERVATION PORTAL**





Lara Lloret Iglesias Instituto de Física de Cantabria

**F** ( A

Instituto de Física de Cantabria

CONSELO SUPERIOR DE INVESTIGACIONES CIENTIFICAS

## Overview

 $\succ$  Open Data at CERN  $\rightarrow$  Why releasing data?

> Data Preservation Policies and goals of the different LHC experiments

CERN Analysis Preservation Portal

► REANA

# Why releasing data?

#### Inclusiveness:

Science should be **open** and **inclusive** 

#### • Engagement:

Being able to access data and *play* with it **engages** people with public research

#### • Education:

Teaching **students** about particle physics and data analysis in general

#### • Society:

A way to give something back to society (that funds our research  $\odot$ )

Scientific Research on Open Data

Enhance scientific output by allowing non-collaboration members to access original data

## How well are we doing right now?

- High Energy Physics is doing well with immediate metadata, such as: beam conditions, event and run numbers, provenance information (processing and reconstruction chain, software versions) recorded together with data at time of data set creation.
- > **Doing worse on** *context* **metadata**, such as : how to pick up the right objects in the data and their documentation, how to know if there are additional selections, corrections...
- The information is readily available and even obvious at time of immediate data analysis, but then easily forgotten

**Open Data and Analysis Preservation help/force us to meet this challenge** Information must be collected and released together with the data

# CERN open data portal **opendata.cern.ch**

> Access point to **growing range of data** produced through research at CERN.

- Disseminates preserved output from various research activities, including accompanying software and documentation needed to understand and analyze the data being shared.
- > Adheres to established global standards in data preservation and Open Science
- The products are shared under **open licenses** and issued with a digital object identifier (DOI) to make them citable.



Close collaboration between experiments, CERN IT and scientific information services
 Many research and educational applications

## Data access policy

ATLAS: <u>http://opendata.cern.ch/record/413</u>
LHCb: <u>http://opendata.cern.ch/record/410</u> → 50% after 5 years
ALICE: <u>http://opendata.cern.ch/record/412</u> → 10% after 5 years
"New" CMS policy data (20th of April 2018): 10.7483/OPENDATA.CMS.7347.JDWH

Release **up to 100% of collision** data within 10 years of data taking (in addition to the current practice of releasing 50% after 3 years)

CMS is the only experiment that has actually released research level data: this currently covers 50% of the Run I data, in line with the CMS policy.

## Data access policy

**Open Data** levels defined in the Data Access Policies:

- ► Level 1: Public results → Papers, HepDATA, etc...
  - Working on providing capability for reinterpretation of searches
- Level 2: Outreach and Education and simplified data formats
  - Lightweight environments to allow the easy exploration of these data
  - Full complexity may also be provided for educational purposes.
- ► Level 3: Reconstructed Data → What is used by analysts
  - Embargo period allowing the collaboration members to perform the analyses
- ► Level 4: Raw Data → What the detector provides
  - Some samples could be released to the interest of data science (e.g. ML and AI)

## The CERN analysis preservation portal (CAP)

- The CERN analysis preservation pursuits the following main goals:
- Describe and structure the knowledge behind a physics analysis aiming for its future reuse
  - Describe all the assets of an analysis and track data provenance.
  - Ensure sufficient documentation and capture associated links.



- Store information about the analysis input data, the analysis code and its dependencies, the runtime computational environment and the analysis workflow steps, and any other necessary dependencies in a trusted digital repository.
- Create a web portal with a **friendly structure** to both register and search the information

## The CERN analysis preservation portal (CAP)

- CAP is built on the Invenio digital library framework
- (used in INSPIRE, CERN Open Data, Zenodo and many others)
- Data are modelled in JSON format
- **JSON Schema** with standard metadata requirements, informed by existing databases and practices in the collaboration
- Elasticsearch cluster for indexing and information retrieval needs
- Open Archival Information System (OAIS) practices to ensure long-term preservation

## Vanilla Mode

Full reproducibility mode- please turn this mode on if you want to capture additional information about main and auxiliary measurements, systematic uncertainties, background estimates, final state particles

Basic Information Please provide some information relevant for all parts of the Analysis here	$\triangleright$	
Information from CADI database CADI info	$\triangleright$	
Input Data Please list all datasets and triggers relevant for your analysis here	$\triangleright$	
N-tuples Production [0 items]	$\triangleright$	
Additional Resources Please provide information about the additional resources of the analysis		
Statistical Treatment	$\triangleright$	

Two different analysis preservation modes:

- Vanilla mode: Basic assets preservation
- **Reuse mode:** Great level of detail  $\rightarrow$  Aiming for reuse

## **Re-Use Mode**

Full reproducibility mode- please turn this mode on if you want to capture additional information about main and auxiliary measurements, systematic uncertainties, background estimates, final state particles			
Basic Information Please provide some information relevant for all parts of the Analysis here	$\triangleright$		
Information from CADI database CADI info	$\triangleright$		
Input Data Please list all datasets and triggers relevant for your analysis here	$\triangleright$		
N-tuples Production [0 items]	$\triangleright$		
Auxiliary Measurements <i>[0 items]</i>	$\triangleright$		
Background Estimation [O items]			
Final Results Please provide information necessary to generate final plots and tables for your analysis.			
Main Measurements Workflows <i>[0 items]</i>	$\triangleright$		
Systematic Uncertainties <i>[0 items]</i>	$\triangleright$		
Additional Resources	$\triangleright$		

Please provide information about the additional resources of the analysis

### **SOME LINKS**

Access to the CERN Analysis Preservation service here: https://analysispreservation.cern.ch

Intructions about how to submit n-tuples and output macros here: https://cernanalysispreservation.readthedocs.io/en/latest/tutorials.html

Submission and retrieval of analysis material can be automatised via the command-line client: https://cap-client.readthedocs.io/

People are already testing it with very positive feedback

## Indexed - searchable

16 results		<pre>&gt; Page1of2</pre>
STATUS draft TYPE cms-analysis-v0.0.1 PHYSICS_OBJECTS jet muon PFMuon GlobalMuon TrackerMuon electron photon MET tau track vertex	16 JME-10-004	We present the results of a visual scan of high E T events (It E T > 60 GeV DR pF E T > 60 GeV) in a large inclusive sample of 7 TeV pp collision data, after applying the official noise clean-up available in CMSSW 3 7 0 patch2. The scan is performed separately for events with tt E T > 60 GeV and pF E T > 60 GeV since two different noise cleaning algorithms are employed. The CMS software Fireworks and PFRootEvent have been used to produce the event displays. The high E T events have been visually inspected and classified in different categories. The results of this scan can provide hints to further improve the noise cleaning and to identify possible problems and inconsistencies in the algorithms employed in CMS for the E T reconstruction.
	FWD-10-005	First measurement is reported of the exclusive two-photon production of muon pairs, $pp \rightarrow p\mu^{\mu}\mu^{-}p$ , in proton-proton collisions at $v_{0} \approx 7$ TeV. For the muon pairs with invariant mass above 11.5 GeV/c 2 and with $pT(\mu) > 4$ GeV/c and $ \eta(\mu)  < 2.1$ , 148 candidates are found in the CMS data sample of 40 pb -1. The characteristic distributions of the muon pairs produced via $\gamma\gamma$ fusion as of the muon acoptanarity and of the pair invariant mass and transverse momentum, are well described by the full Monte Carlo simulation using the LPAR event generator. Small and well understood background to the process is observed, and it is shown that $pp \rightarrow p\mu^{\mu}\mu^{-}p$ provides a reliable absolute normalization of the LHC luminosity.
	4 4 10 6 2 2 ELECTRON MUON	Previous measurements in ep and hadron-hadron colliders demonstrated that events with large rapidity gaps (LRP) can be described by diffractive interactions. A quanti- tative interpretation of LRG events at hadron-hadron colliders is complicated by the fact that the LRG signal is destroyed or diminished by multiparton interactions. In this paper we study the correlations of the energy flow in the forward detectors and the track multiplicity in the central detector using events with centrally produced W and 2 bosons, identified with their leptonic decays. The analysis uses the entry of TeV pp collision high quality data sample, about 36pb -1, recorded during the 2010 LHC operation and strict conditions for single pp vertex events. The observed forward energy deposits, their correlations and the track multiplicities in the central part of the CMS experiment are compared with different Monte Carlo
	2 AN-2011/062	Yields of prompt and non-prompt J/u, as well as Y(1S) mesons, are measured by the CMS experiment via their µ+ µ- decays in PbPb and pp collisions at visNN = 2.76 TeV for quarkonium rapidity (y) < 2.4. Differential cross sections and nuclear modification factors are reported as functions of y and transverse momentum pT, as well as collision centrality. For prompt J/µ with relatively high pT (6.5 < pT < 30 GeV/c), a strong, centrality-dependent suppression is observed in PbPb collisions, compared to the yield in pp collisions scaled by the number of inelastic nucleon-nucleon collisions. In the same kinematic range, a suppression of non-prompt J/µ, which is sensitive to the in-medium b-quark energy face, is measured for the first time. Also the low-pT Y(1S) mesons are suppressed in PbPb collisions.
	AN-2011/103	Previous measurements in ep and hadron-hadron colliders demonstrated that events with large rapidity gaps (LRP) can be described by diffractive interactions. A quanti-tative interpretation of LRG events at hadron-hadron colliders is complicated by the fact that the LRG signal is destroyed or diminished by multiparton interactions. In this paper we study the correlations of the energy flow in the forward detectors and the track multiplicity in the central detector using events with centrally produced W and 2 bosons, identified with their leptonic decays. The analysis uses the entire 7 TeV pp collision high quality data sample, about 36pb ~1, recorded during the 2010 LNC operation and strict conditions for single pp vertex events. The observed forward energy deposits, their correlations and the track multiplicities in the central part of the CMS experiment are compared with different Monte Carlo
	AN-2010/411 ELECTRON MUON MET	This note describes the search for the Higgs boson 4 in the H $\rightarrow$ WW $\rightarrow$ v v decay channel in about 35.5 pb-1 of pp collision data at s = 7 TeV collected by the CMS detector at the LHC. Event yields are presented along with background predictions, expected signal yields, and the associated uncertainties for a sequential and a multivariate analyses. No excess above the Standard Model predictions is found in the current data sample and limits on the Higgs boson production cross-section times branching ratio are derived. With the current amount of data, the observed limits have no sensitivity to the 5M Higgs boson, but compared to the recent theoretical calculations performed in the context the Standard Model with a four fermion generation, allow for excluding the Higgs boson with a mass in the 144-207 GeV range at 95% confidence level.

## **Command line client**

Command line client to facilitate workflow automation:

- Allows to create/edit JSON files with the analysis information and submit it directly from the terminal
- Information propagated to the web portal

```
$ pip install cap-client
$ export CAP_SERVER_URL=https://analysispreservation.cern.ch/
$ export CAP_ACCESS_TOKEN=<your generated access token from server>
$ cap-client files upload <file path> --pid/-p <existing pid>
$ cap-client files upload file.json -p 89b593c498874ec8bcafc88944c458a7
File uploaded successfully.
```



#### a system for reusable analysis execution on the cloud

O https://reanahub.io

#### supporting multiple scenarios

- multiple computing clouds  $\rightarrow$  CERN OpenStack
- multiple running environments  $\rightarrow$  Docker with CVMFS
- multiple resource orchestration  $\rightarrow$  Kubernetes
- multiple workflow engines  $\rightarrow$  Yadage
- multiple shared storage systems  $\rightarrow$  Ceph, EOS





## Conclusions

- All four experiments are doing as much as they can to make the Open Data goals advance → Limited by experiment resources
- Open Data enables us to engage with the outside world on a more meaningful level enabling us to show how we work
- There are plenty of tools and datasets ready for you to use
- The data can be used to **develop lab courses, full visualisations, create teaching** materials...and also research results.
- Working on **analysis preservation** and **reproducibility** 
  - CERN Analysis Preservation Portal (CAP) already being filled
- REANA: Platform for reproducible research analysis based on dockers is in place
  - Next steps: Integration with CAP

## FURTHER READING



Perspective Open Access Published: 15 November 2018

## Open is not enough

Xiaoli Chen, Sünje Dallmeier-Tiessen <sup>™</sup>, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, Diego Rodriguez Rodriguez, Tibor Šimko <sup>™</sup>, Tim Smith, Ana Trisovic <sup>™</sup>, Anna Trzcinska, Ioannis Tsanaktsidis, Markus Zimmermann, Kyle Cranmer, Lukas Heinrich, Gordon Watts, Michael Hildreth, Lara Lloret Iglesias, Kati Lassila-Perini & Sebastian Neubert

Nature Physics 15, 113–119 (2019)| Download Citation ±9707 Accesses5 Citations151 AltmetricMetrics ≫

# Backup

#### The other experiments

#### ALICE APPROACH:

- ALICE LEGO trains: Centralized analysis system within ALICE to run analyses on the Grid
- $\succ$  Only 5 scripts  $\rightarrow$  easy to automatize and preserve

#### ATLAS APPROACH:

- ATLAS has a already had quite extensive analysis tracking tool: "Analysis Glance" — recently expanded quiet a bit
- > Used to track analysis through the entire lifecycle
- Large amount of Metadata
- ➤ Envisioned to be main input source to CAP → Need work to build ingestion route Glance → CAP

#### LHCb APPROACH:

- > Two approaches:
  - Partial Reproducibility: Store ntuples and anything happening afterwards
  - Total Reproducibility: Needs full workflow and adecuate framework (docker)
- Different pipeline tools under evaluation