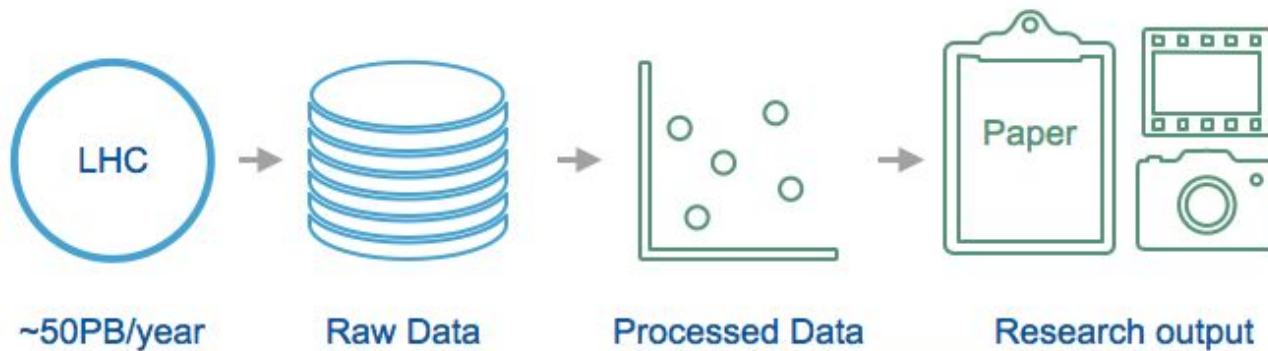




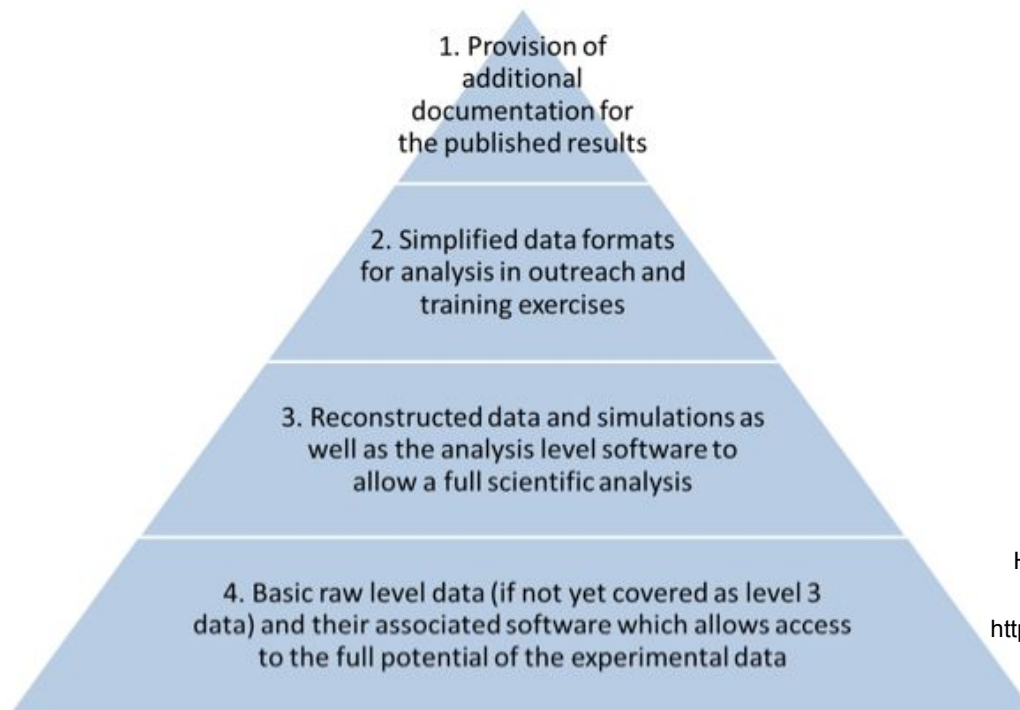
# Publishing, preservation and linking services at CERN

In the context of the LODES proposal

# Data in HEP



# Data in HEP



Herterich, P., & Dallmeier-Tiessen, S. (2016). Data Citation Services in the High-Energy Physics Community. *D-Lib Magazine*, 22(1/2).  
<http://doi.org/10.1045/january2016-herterich>

# Access to HEP data: Data Policies

## LHCb External Data Access Policy

### ALICE data preservation strategy

Sunday, October 6, 2013

The data harvested by the ALICE investment in human and financial resources provides information for the in depth understanding of the collision system.

Approved CB 20<sup>th</sup> June 2014

### ATLAS Data Access Policy

May 21<sup>st</sup> 2014

### 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<sup>1</sup>. 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.

publication policy. This is to data at different levels as take the data available in a

rate document. The some requirements with making | resources will be required to

on can be considered at four conditions see Def [1]. This

- ❑ Open data for three levels
- ❑ Immediate release for level 1
- ❑ Embargo periods for level 3
- ❑ CC0 – public domain dedication



# Services

- ❑ **CERN Open Data**: platform for curated releases of CERN data sets, software, supplementary materials, etc., over 1PB of data
- ❑ **CERN Analysis Preservation**: restricted-access service, preservation of knowledge and assets of physics analyses
- ❑ **REANA (REusable ANALysis)**: platform for running research data analyses on containerised compute clouds
- ❑ **HEPData**: repository for tabular data associated with publications
- ❑ **INSPIRE-HEP**: core HEP literature aggregator, “long tail” of data

# Metadata granularity - example 1: CAP



## Basic Information

**Analysis Name** - Provide a name for your analysis. This will be displayed as an analysis title when shared.

**Measurement** - Provide a Measurement type. This will be displayed as an analysis title when shared.

Proponents

Status

Reviewers

Review eGroup

Institutes Involved

Keywords

## Input Data

Primary Datasets

Monte Carlo Signal Datasets

Monte Carlo Background Datasets

Triggers

Official JSON files

## Analysis Software

N-tuples Production [0 items]

Main Measurements Workflows [0 items]

Auxiliary Measurements [0 items]

Background Estimation [0 items]

Systematic Uncertainties [0 items]

Final Results

Please provide information necessary to generate final plots and tables for your analysis.

# Metadata granularity - example 2: COD



```
"accelerator": "CERN-LHC",
"collaboration": {
  "name": "CMS collaboration",
  "recid": "451"
},
"collision_information": {
  "energy": "8TeV",
  "type": "pp"
},
"distribution": {
  "formats": [
    "aodsim",
    "root"
  ],
  "number_events": 21732511,
  "number_files": 2090,
  "size": 6211342356850
},
"pileup": {
  "description": "<p>To make these simulated data comparable with the collision data, <a href=\"/about/CMS-Pileup-Simulation\">pile-up events</a> are added to the simulated event in this step.</p>"
},
"relations": [
  {
    "doi": "10.7483/OPENDATA.CMS.PDY4.7H2H",
    "recid": "4",
    "title": "/Electron/Run2010B-Apr21ReReco-v1/AOD",
    "type": "isChildOf"
  }
]
"run_period": "2012A-2012D",
"system_details": {
  "global_tag": "START53_V27",
  "release": "CMSSW_5_3_32"
},
"use_with": {
  "description": "The event displays prepared for the Masterclasses W-Path."
},
"validation": {
  "description": "The generation and simulation of simulated Monte Carlo data has been validated through general CMS validation procedures."
}
```



# Interoperability

Established practices for “FAIR” data and beyond:

- ❑ Persistent Identifiers
- ❑ High-quality metadata specific to individual collaborations/working groups
- ❑ Open-source software and tools (Invenio)
- ❑ Data & software citation
- ❑ JSON export & schemata, schema.org markup using JSON-LD
- ❑ Reuse
  - ❑ E.g. capturing all the elements needed to understand and rerun an analysis even several years later (data, software, environment, workflow, context, documentation) and linking them together persistently





# To conclude

## **Make it your own:**

HEP data is highly complex, HEP repositories don't have to be

- Flexible open source software
- Standards that are not restrictive and can be adapted
- Tailored Open Science practices

## URLs

[opendata.cern.ch](https://opendata.cern.ch)

[github.com/cernanalysispreservation/analysispreservation.cern.ch](https://github.com/cernanalysispreservation/analysispreservation.cern.ch)

[reana.io](https://reana.io)

[hepdata.net](https://hepdata.net)

[inspirehep.net](https://inspirehep.net)



**CERN IT** H. Hirvonsalo, D. Kousidis, D. Rodriguez, T. Šimko · **CERN SIS** S. Dallmeier-Tiessen, S. Feger, P. Fokianos, A. Lavasa, I. Tsanaktsidis, A. Trisovic, A. Trzcinska · **ALICE** D. Berzano, M. Gheata, C. Grigoras, Y. Schutz, M. Zimmermann · **ATLAS** J. Berlingen, K. Cranmer, L. Heinrich, L. Henkelmann, A. Sanchez Pineda, D. Rousseau, F. Socher · **CMS** A. Calderon, E. Carrera, A. Geiser, A. Huffman, C. Lange, K. Lassila-Perini, L. Lloret, T. McCauley, A. Rao, A. Rodriguez Marrero · **LHCb** S. Amerio, C. Burr, B. Couturier, S. Neubert, C. Parkes, A. Pearce, S. Roiser · **DASPOS** M. Hildreth · **DPHEP** J. Shiers