

# CERN (OPEN) DATA SERVICES

---

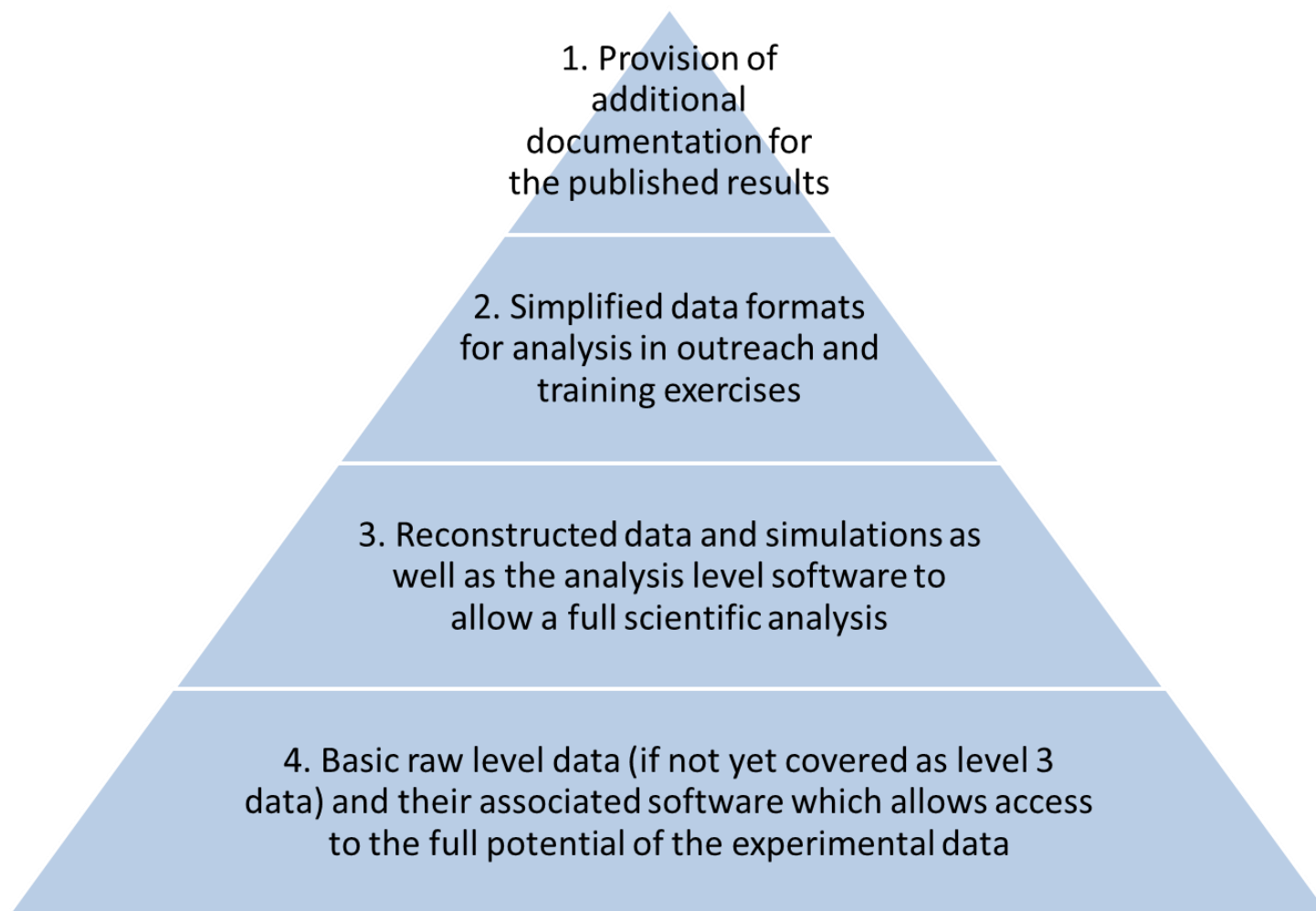
Invenio User Group Workshop 2015

13<sup>th</sup> October 2015

Patricia Herterich

CERN GS-SIS-OA & Humboldt-Universität zu Berlin

# Data in High-Energy Physics



# CERN Open Data Portal

- LHC collaborations' data policies\*

“[...] Data with high abstraction, such as AOD, will be conditionally made publicly available after an embargo period of 5 years after publication for 10% of the data and 10 years for 100% of the data [...]”

—ALICE Data Policy

## But where should the data be published?

\* Available at <http://opendata.cern.ch/collection/Data-Policies>

# CERN Open Data Portal

- Released Nov 2014
- <http://opendata.cern.ch/>
- Based on Invenio 2.0
- <https://github.com/cernopendata>
- Audience:
  - data miners
  - citizen scientists
  - high-school students
  - general public
  - But also physicists

## Education

Visualise events, check reconstructed data, run tools or build your own!

Start learning

## Research

Get the genuine working environments, virtual machines and datasets to start your research

Start analysing

### Education



The CMS (Compact Muon Solenoid) experiment is one of two large general-purpose detectors built on the Large Hadron Collider (LHC). Its goal is to investigate a wide range of physics such as the characteristics of the Higgs boson, extra dimensions or dark matter.

Explore CMS >



ALICE (A Large Ion Collider Experiment) is a heavy-ion detector designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms. More than 1000 scientists are part of the collaboration.

Explore ALICE >



The ATLAS (A Toroidal LHC Apparatus) experiment is a general purpose detector exploring topics like the properties of the Higgs-like particle, extra dimensions of space, unification of fundamental forces, and evidence for dark matter candidates in the Universe.

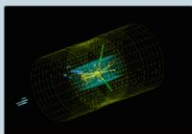
Explore ATLAS >



The LHCb (Large Hadron Collider beauty) experiment aims to record the decay of particles containing b and anti-b quarks, known as B mesons. The detector is designed to gather information about the identity, trajectory, momentum and energy of each particle.

Explore LHCb >

For education purposes, the complex primary data need to be processed into a format (examples below) that is good for simple applications. Get in touch if you wish to build your own applications similar to those shown here



Visualise events >



Visualise histograms >



Learning Resources >

### Research



To analyse CMS data, a Virtual Machine with the CMS analysis environment is provided. The data can be accessed directly through the VM. In the primary datasets, no selection nor identification criteria have been applied. For this release, no simulated Monte Carlo datasets are provided.

Explore CMS >



According to the ALICE data preservation strategy, reconstructed data and Monte Carlo data as well as the analysis software and documentation needed to process them will be made available on a time scale of 5 years (for 10% of the data). Thus, the first release of ALICE research data will happen in 2018.



According to the ATLAS Data Access Policy, reconstructed data and accompanying tools will be released after reasonable embargo periods.



According to the LHCb External Data Access Policy, reconstructed data and accompanying tools will be released after reasonable embargo periods.

For research purposes, specific software environments and tools need to be deployed to analyse these complex primary data. In addition to the data below, you will find instructions for setting up your working environments here



Install your Virtual Machine >



Start analysing the data >

# COD – Event visualisation

opendata COM

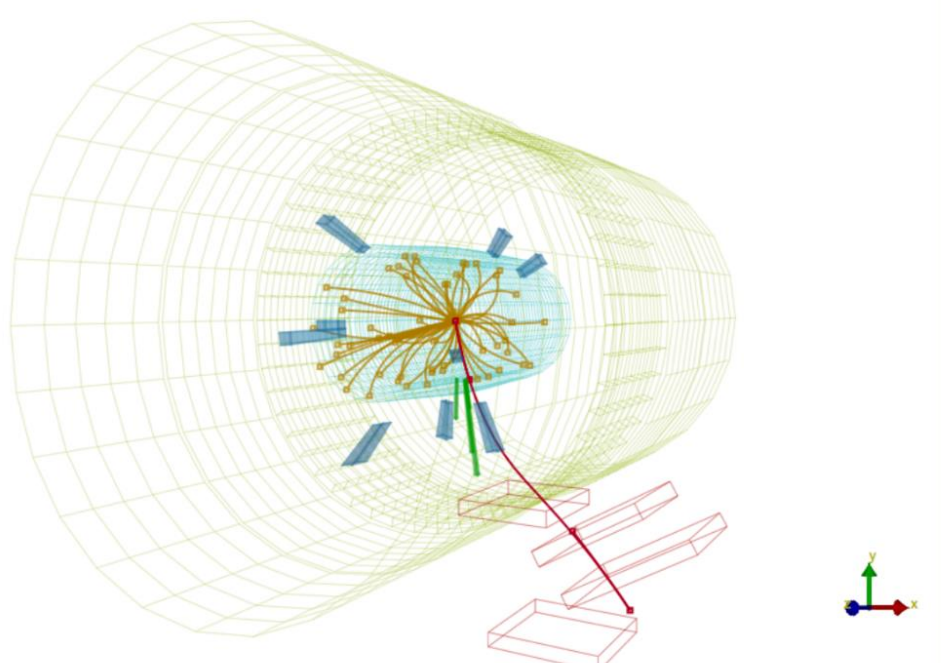
ABOUT SEARCH EDUCATION RESEARCH

Home > Education > Visualise Events > CMS

## Explore CMS open data and visualise detector events

Need HELP?

/Mu.Jg:Events/Run\_146436/Event\_90626440



Detector Model	
Tracker Barrels	<input type="checkbox"/>
Tracker Endcaps	<input type="checkbox"/>
ECAL Barrel	<input checked="" type="checkbox"/>
ECAL Endcaps	<input type="checkbox"/>
ECAL Preshower	<input type="checkbox"/>
HCAL Barrel	<input type="checkbox"/>
HCAL Endcaps	<input type="checkbox"/>
HCAL Outer	<input checked="" type="checkbox"/>
HCAL Forward	<input type="checkbox"/>
Drift Tubes (muon)	<input type="checkbox"/>
Cathode Strip Chambers (muon)	<input type="checkbox"/>
Resistive Plate Chambers (muon)	<input type="checkbox"/>
<b>Tracking</b>	<input checked="" type="checkbox"/>
Tracks (reco.)	<input checked="" type="checkbox"/>
<b>ECAL</b>	<input checked="" type="checkbox"/>
Barrel Rec. Hits	<input checked="" type="checkbox"/>
Endcap Rec. Hits	<input type="checkbox"/>
Preshower Rec. Hits	<input type="checkbox"/>
<b>HCAL</b>	<input checked="" type="checkbox"/>
Barrel Rec. Hits	<input checked="" type="checkbox"/>
Endcap Rec. Hits	<input checked="" type="checkbox"/>
Forward Rec. Hits	<input checked="" type="checkbox"/>
Outer Rec. Hits	<input type="checkbox"/>
<b>Muon</b>	<input checked="" type="checkbox"/>
Matching muon chambers	<input checked="" type="checkbox"/>
<b>Physics Objects</b>	<input checked="" type="checkbox"/>
Electron Tracks (GSF)	<input checked="" type="checkbox"/>
Tracker Muons (Reco)	<input checked="" type="checkbox"/>
Stand-alone Muons (Reco)	<input checked="" type="checkbox"/>
Global Muons (Reco)	<input checked="" type="checkbox"/>
Calorimeter Energy Towers	<input type="checkbox"/>
Jets	<input type="checkbox"/>
Missing Et (Reco)	<input type="checkbox"/>

# CMS Primary Datasets

Photon primary dataset in AOD format from RunB of 2010 (/Photon/Run2010B-Apr21Reco-v1/AOD) <sup>2014</sup>

/Photon/Run2010B-Apr21Reco-v1/AOD  
CMS collaboration

**Cite as:** CMS collaboration (2014). Photon primary dataset in AOD format from RunB of 2010 (/Photon/Run2010B-Apr21Reco-v1/AOD). [10.7483/OPENDATA.CMS.OKAX.PSW6](https://doi.org/10.7483/OPENDATA.CMS.OKAX.PSW6)

Collection CMS Primary Datasets Collision Energy 7TeV Accelerator CERN-LHC Experiment CMS

## Description

Photon primary dataset in AOD format from RunB of 2010

## Characteristics

Dataset: 25465895 events 2814 files 2.6 TB in total

## System Details

Software release: CMSSW\_4\_2\_1\_patch1

## Indexes

CMS\_Run2010B\_Photon\_AOD\_Apr21Reco-v1\_0002\_file\_index.txt Size: 41.8 kB  
Description: Photon AOD dataset file index (3 of 6) for access to data via CMS virtual machine

CMS\_Run2010B\_Photon\_AOD\_Apr21Reco-v1\_0001\_file\_index.txt Size: 55.8 kB  
Description: Photon AOD dataset file index (2 of 6) for access to data via CMS virtual machine

CMS\_Run2010B\_Photon\_AOD\_Apr21Reco-v1\_0004\_file\_index.txt Size: 46.6 kB  
Description: Photon AOD dataset file index (5 of 6) for access to data via CMS virtual machine

CMS\_Run2010B\_Photon\_AOD\_Apr21Reco-v1\_0003\_file\_index.txt Size: 77.2 kB  
Description: Photon AOD dataset file index (4 of 6) for access to data via CMS virtual machine

CMS\_Run2010B\_Photon\_AOD\_Apr21Reco-v1\_0005\_file\_index.txt Size: 35.6 kB  
Description: Photon AOD dataset file index (6 of 6) for access to data via CMS virtual machine

CMS\_Run2010B\_Photon\_AOD\_Apr21Reco-v1\_0000\_file\_index.txt Size: 94.8 kB  
Description: Photon AOD dataset file index (1 of 6) for access to data via CMS virtual machine

0072FAED-2471-E011-87D2-0018FE2930C6.root: xrootd

Size: 527.0 MB

Download

Files 1 - 5 out of 2814

## How were these data selected?

Events stored in this primary dataset were selected because of presence of at least one high-energy photon in the event.

## How were these data validated?

During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and physics objects (tracking, electron, muon, gamma, jet and MET) show the expected performance. Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:

- CMS data quality monitoring: Systems and experiences
- The CMS Data Quality Monitoring software experience and future improvements
- The CMS data quality monitoring software: experience and future prospects

## How can you use these data?

You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in

- How to install the CMS Virtual Machine
- Getting started with CMS open data

## Issues & Limitations

This dataset contains all runs from 2010 RunB. The list of validated runs, which must be applied to all analyses, can be found in

CMS list of validated runs Cert\_136033-149442\_7TeV\_Apr21Reco\_Collisions10\_JSON\_v2.txt

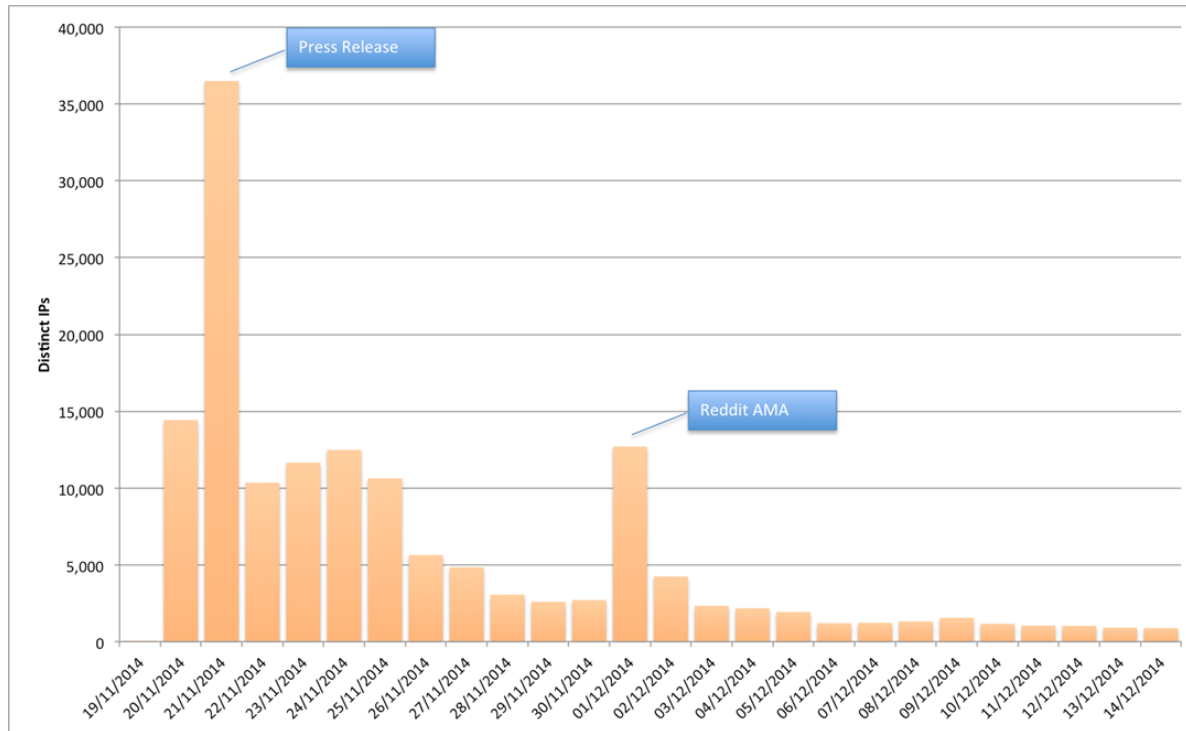
## Disclaimer

The open data are released under the [Creative Commons CC0 waiver](#). Neither CMS nor CERN endorse any works, scientific or otherwise, produced using these data. All releases will have a unique DOI that you are requested to cite in any applications or publications.



Export MARXML

# Impact



- steady-state numbers after 14/12/2014:
    - ~1000 visitors, out of which about
      - ~10 people download EOS files
      - ~400 people look at detailed record pages
- resulting in various amounts GB being served

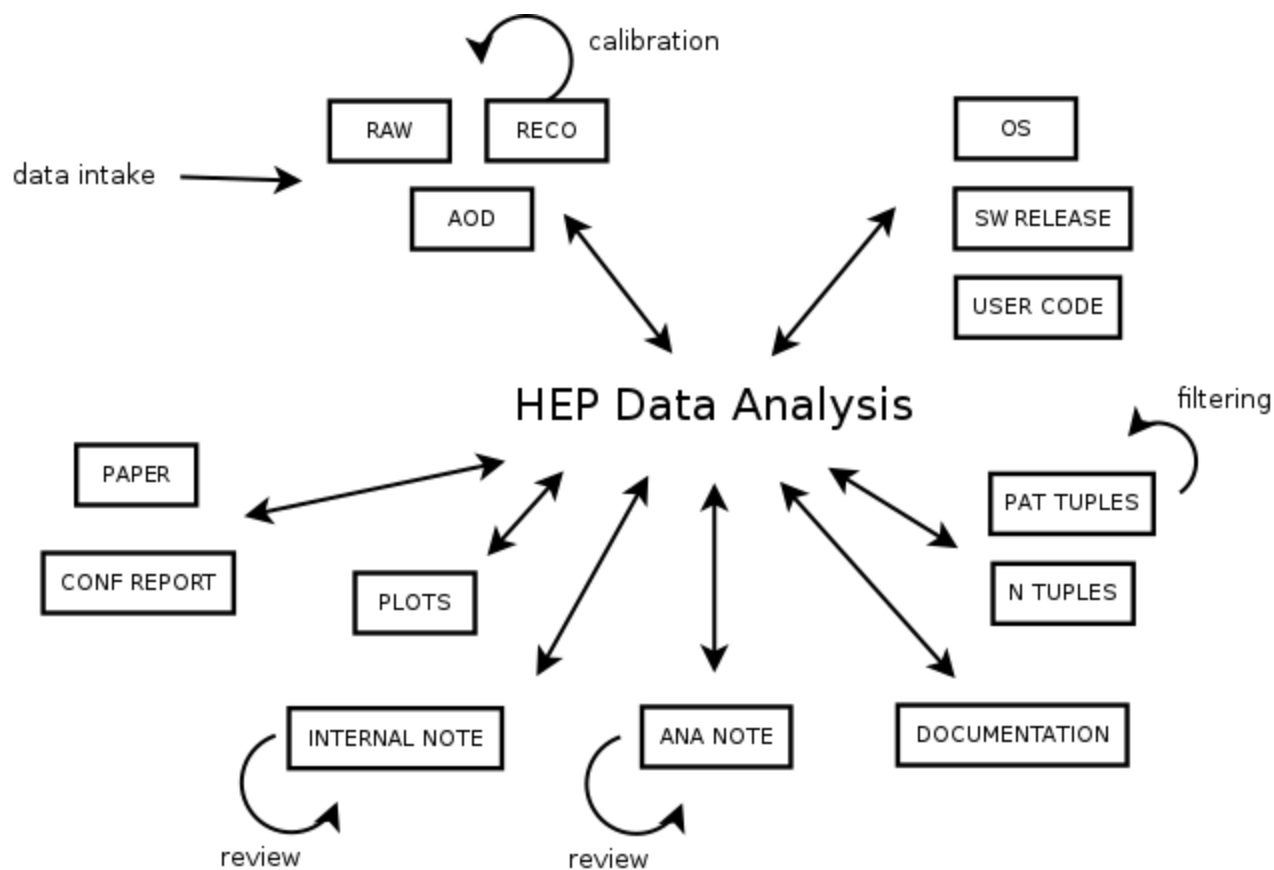


# CERN Analysis Preservation

- “**closed** counterpart” to CERN Open Data that captures the complexity of
  - The data
  - The processing steps
  - Code involved
  - Documentation, Physics information
  - Peer review, QA

i.e. all the information contributing to the research claim/presentation/publication to enable future reuse

# Preserve an analysis?



# Use cases

- Originally: Preservation of data and associated objects and information

Extended to:

- Making user generated data/content discoverable, e.g. link them to primary data [search]
- Facilitate easy intra- or inter collaboration data/method comparisons and validation
- Automated integration of data/physics info into approval workflows

# Pilot Prototype

<https://analysis-preservation.cern.ch/>

CERN Accelerating science

Sign in Directory

## CERN Analysis Preservation

Demo (your data will NOT be preserved!)

Search Deposit

INVENIO Search Deposit Help

### Physics Information

**Primary Data Set**

**MC Data Set Path**

**Trigger Selection**

Upon full implementation, this selection box will cover all years (as this is just sets for 2013) and will be filtered based on the input from the primary data set above

**Final State Particles**

+ Add Final State Particles

Particle - Number - PT Cut - ETA Cut

**Physics Objects**

**Keywords**

+ Add another keyword

**Comments**



### AOD Production Step

**OS**

**Analysis Software**

**User Code**    Harvest  Link only

**Input data files**  AOD Primary Data Sets  Taken from output of previous analysis step

**Output Data Files**   Harvest  Link only

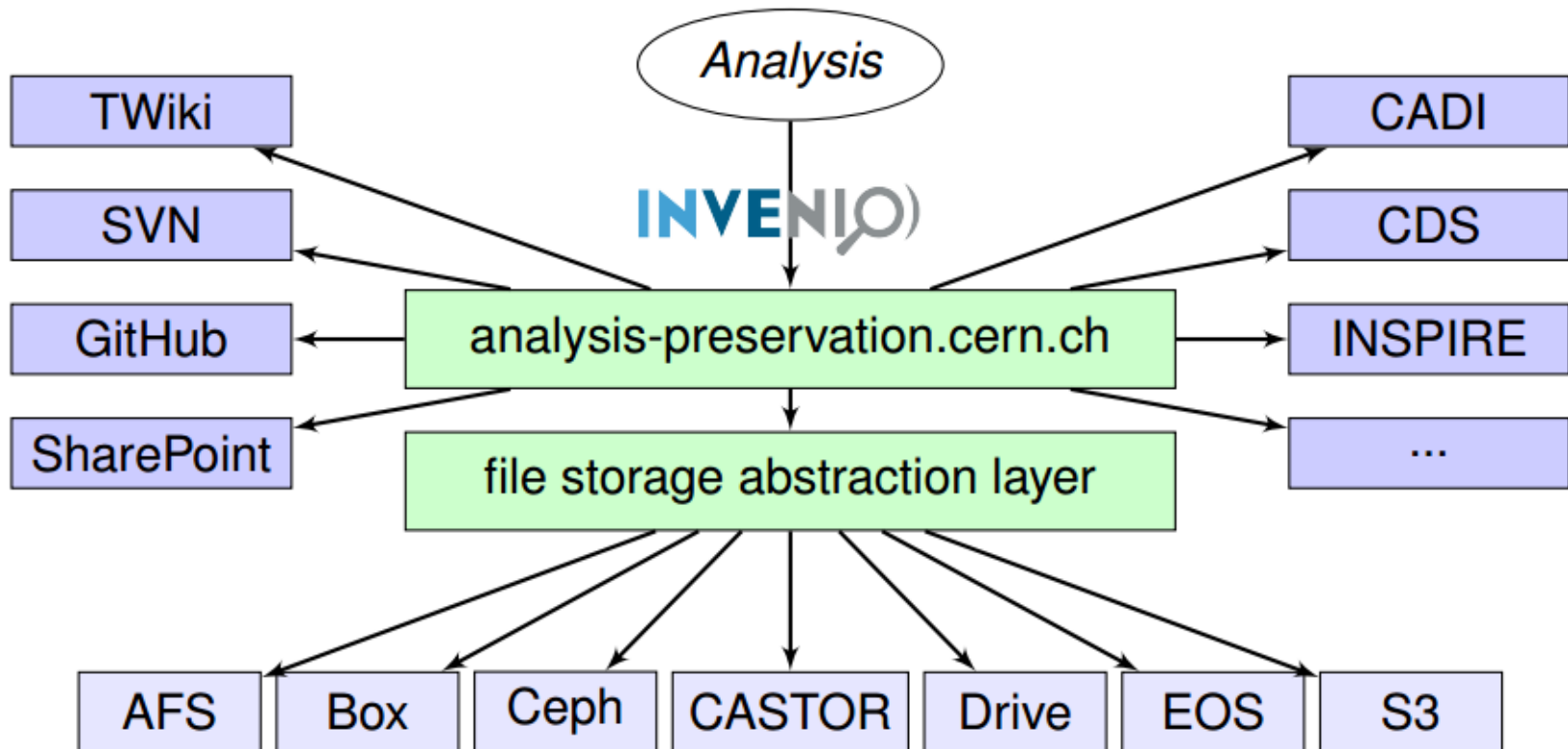
+ Add Output Data Files

**How to reproduce**

**Keywords**

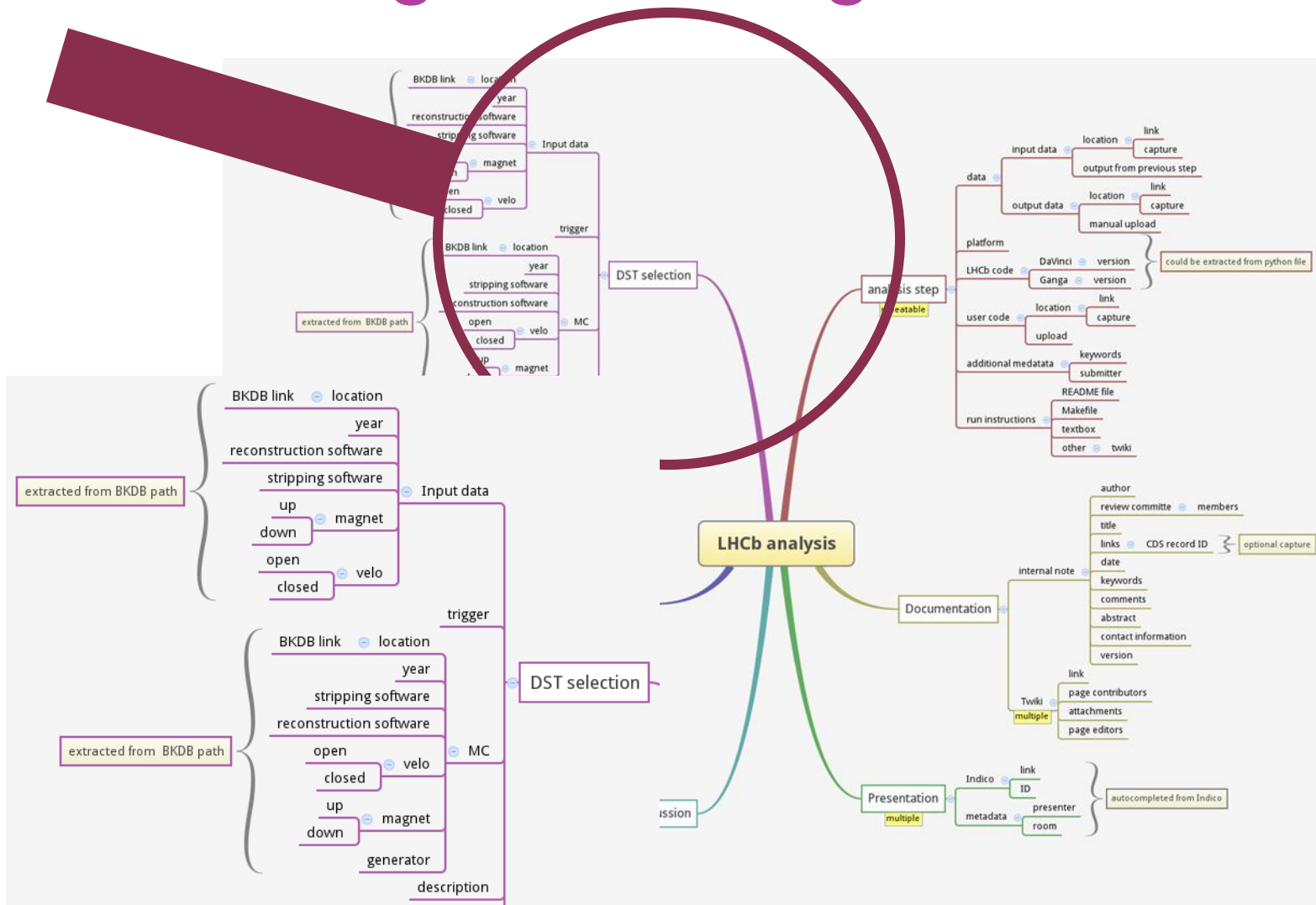
+ Add another keyword

# CAP – system architecture



<https://github.com/cernanalysispreservation>

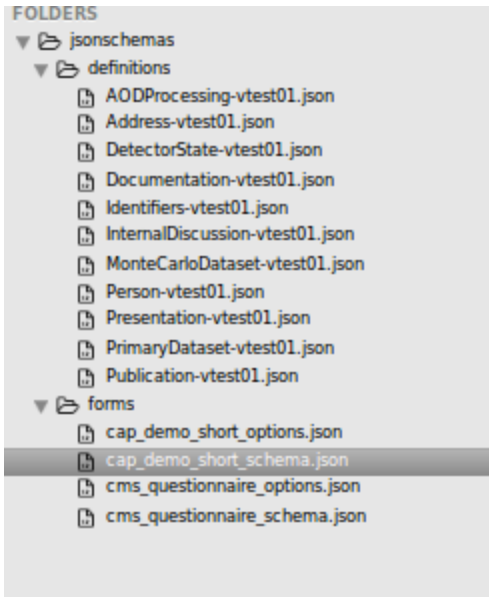
# Knowledge modelling



# JSON based metadata schema

- Allows composition of a metadata schema
  - Re-use/integration of existing schemas
  - Sub-schemas
- Easily extendable to JSON-LD
- CAP schema follows best practices for data description (Data Catalog Vocabulary (DCAT))

# JSON based metadata schema



```
"$schema": "http://json-schema.org/draft-04/schema#",
"title": "CAP Schema Example",
"description": "Schema example for records Schema example for records Schema example for records",
"properties": {
  "common": {
    "title": "CAP Common",
    "type": "object",
    "id": "common",
    "properties": {
      "experiment": {
        "title": "Experiment",
        "description": "Choose an experiment from the list",
        "enum": [
          "ALICE",
          "ATLAS",
          "CMS",
          "LHCb"
        ],
        "type": "string"
      },
      "analysis_title": {
        "description": "Analysis Description",
        "title": "Analysis Title",
        "type": "string"
      },
      "creator": {
        "id": "creator",
        "title": "Creator(s)",
        "items": {
          "title": "Creator",
          "allOf": [
            {
              "$ref": "/jsonschemas/definitions/Person-vtest01.json#"
            },
            {
              "properties": {
                "identifiers": {
```



# JSON based submission forms

Delete Save Submit

## CMS JSON Analysis Demo SHORT

THIS IS JUST A DEMO. DATA IS *NOT* STORED

Access to all submitted data will be restricted to the CMS collaboration only.

### CAP Schema Example

#### CAP Common

Experiment

Choose an experiment from the list

Analysis Title

#### Physics Information

Add detector final states

Detector Final States

Physics Object	<input type="text"/>
Count	<input type="text"/>
eta	<input type="text"/>
pT	<input type="text"/>
Transverse Energy	<input type="text"/>
Charge	<input type="text"/>

# Thanks to

**CERN IT** J. Cowton, J. Delgado, J. Kunčar, M. Neumann, T. Smith, T. Šimko

**CERN SIS** S. Dallmeier-Tiessen, A. Dani, P. Fokianos, L. Rueda

**ALICE** M. Gheata, C. Grigoras

**ATLAS** K. Cranmer, L. Heinrich, D. Rousseau, F. Socher

**CMS** A. Calderon, A. Huffman, K. Lassila-Perini, T. McCauley, A. Rao, A. Rodriguez Marrero

**LHCb** S. Amerio, M. Bettler, B. Couturier, T. Head, A. Trisovic, A. Ustyuzhanin

**CERN CernVM** J. Blomer

**CERN EOS** L. Mascetti

**DASPOS** M. Hildreth, C. Vardeman, G. Watts

**DPHEP** F. Berghaus, J. Shiers

