CERN Open Data

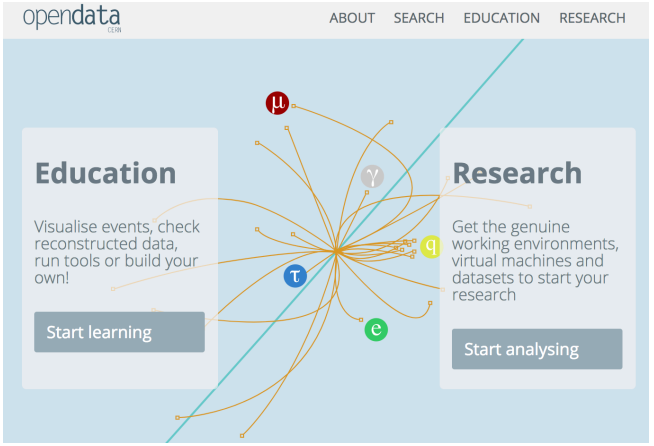
Sünje Dallmeier-Tiessen
for many others in CERN IT and CERN Scientific Information Service

opendata-support@cern.ch

opendata.cern.ch

What is CERN Open Data

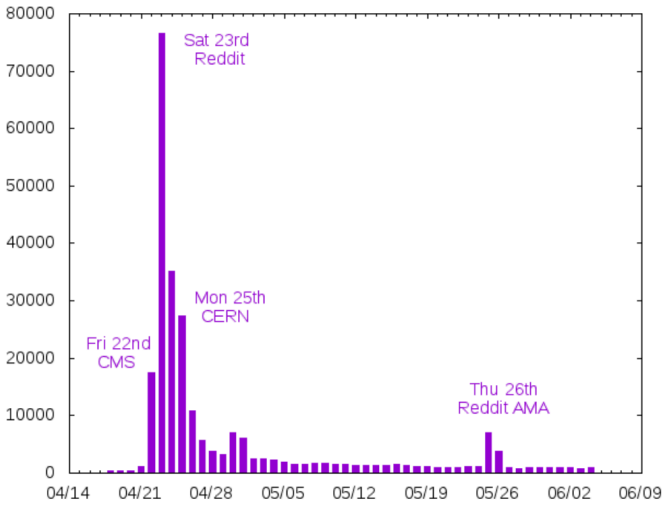
- Curated data releases to the public
- Various sizes and complexities
- For a diverse audience
  - general public and citizen scientists
  - students - high school and university
  - data miners
  - researchers

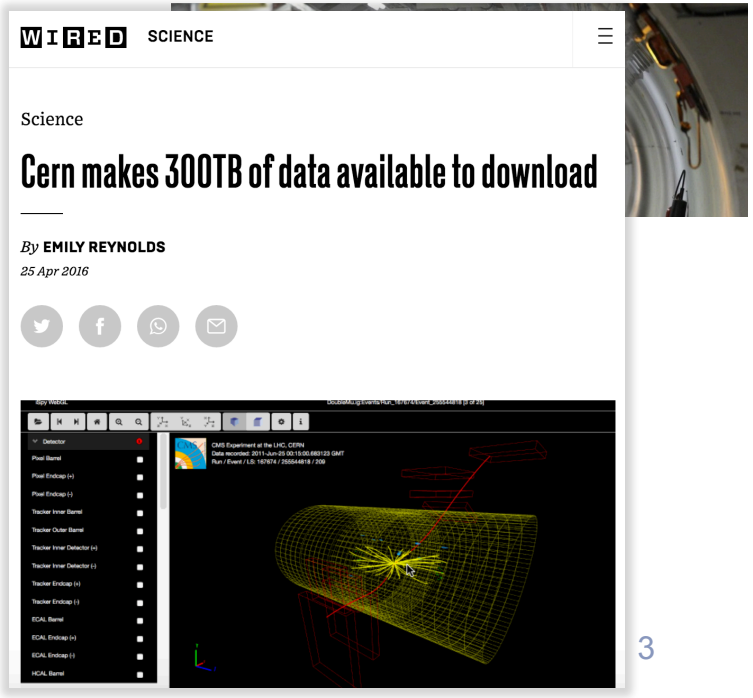


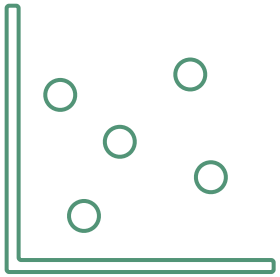
Who cares?

In about a month following CMS 2011 “300TB” open data release (April 2016)

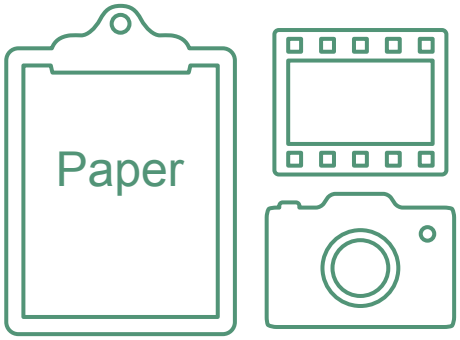
- 210,000 distinct users visited the site
- 66,000 distinct users used event display
- 37,000 distinct users viewed data records
- 3,000 distinct users used histogramming







Processed Data



Research output

CERN ANALYSIS PRESERVATION

http://analysispreservation.cern.ch

REANA

https://github.com/reanahub

CERN OPEN DATA

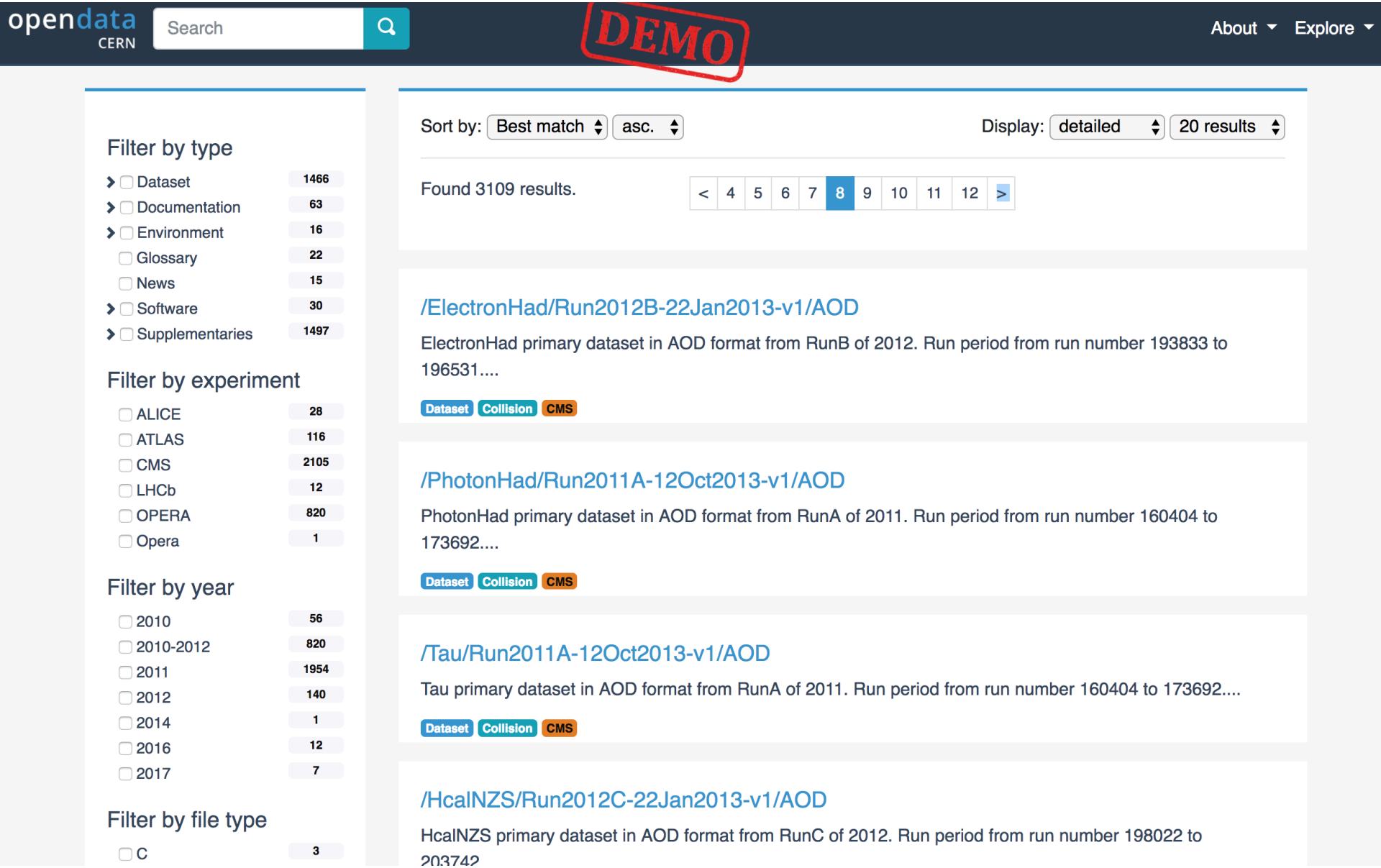
http://opendata.cern.ch

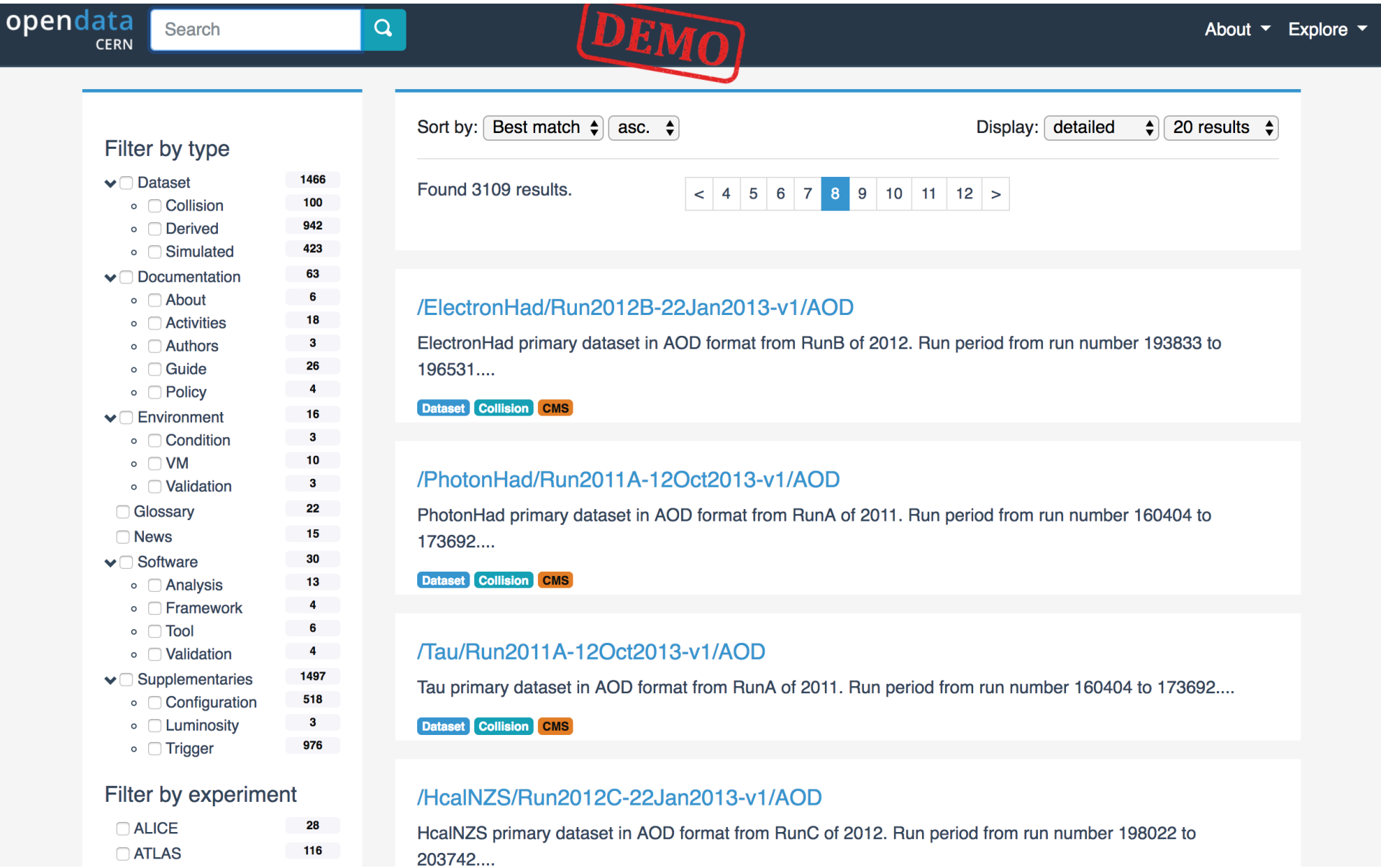
ARXIV/ INSPIRE/ CDS

https://inspirehep.net

New CERN Open Data

BETA demo


opendata
CERN
Search
DEMO
About
Explore
Filter by type
Dataset 1466
Documentation 63
Environment 16
Glossary 22
News 15
Software 30
Supplementaries 1497
Filter by experiment
ALICE 28
ATLAS 116
CMS 2105
LHCb 12
OPERA 820
Opera 1
Filter by year
2010 56
2010-2012 820
2011 1954
2012 140
2014 1
2016 12
2017 7
Filter by file type
C 3
Sort by: Best match asc.
Display: detailed 20 results
Found 3109 results.
< 4 5 6 7 8 9 10 11 12 >
/ElectronHad/Run2012B-22Jan2013-v1/AOD
ElectronHad primary dataset in AOD format from RunB of 2012. Run period from run number 193833 to 196531....
Dataset Collision CMS
/PhotonHad/Run2011A-12Oct2013-v1/AOD
PhotonHad primary dataset in AOD format from RunA of 2011. Run period from run number 160404 to 173692....
Dataset Collision CMS
/Tau/Run2011A-12Oct2013-v1/AOD
Tau primary dataset in AOD format from RunA of 2011. Run period from run number 160404 to 173692....
Dataset Collision CMS
/HcalNZS/Run2012C-22Jan2013-v1/AOD
HcalNZS primary dataset in AOD format from RunC of 2012. Run period from run number 198022 to 203742


opendata
CERN
Search
DEMO
About
Explore
Filter by type
Dataset 1466
Collision 100
Derived 942
Simulated 423
Documentation 63
About 6
Activities 18
Authors 3
Guide 26
Policy 4
Environment 16
Condition 3
VM 10
Validation 3
Glossary 22
News 15
Software 30
Analysis 13
Framework 4
Tool 6
Validation 4
Supplementaries 1497
Configuration 518
Luminosity 3
Trigger 976
Filter by experiment
ALICE 28
ATLAS 116
Sort by: Best match asc.
Display: detailed 20 results
Found 3109 results.
< 4 5 6 7 8 9 10 11 12 >
/ElectronHad/Run2012B-22Jan2013-v1/AOD
ElectronHad primary dataset in AOD format from RunB of 2012. Run period from run number 193833 to 196531....
Dataset Collision CMS
/PhotonHad/Run2011A-12Oct2013-v1/AOD
PhotonHad primary dataset in AOD format from RunA of 2011. Run period from run number 160404 to 173692....
Dataset Collision CMS
/Tau/Run2011A-12Oct2013-v1/AOD
Tau primary dataset in AOD format from RunA of 2011. Run period from run number 160404 to 173692....
Dataset Collision CMS
/HcalNZS/Run2012C-22Jan2013-v1/AOD
HcalNZS primary dataset in AOD format from RunC of 2012. Run period from run number 198022 to 203742....

BTag primary dataset in AOD format from Run of 2012 (/BTag/Run2012C-22Jan2013-v1/AOD) 2017

/BTag/Run2012C-22Jan2013-v1/AOD CMS collaboration

Dataset Collision Collision Energy: 8TeV Experiment: CMS Accelerator: CERN-LHC

Description

BTag primary dataset in AOD format from RunC of 2012. Run period from run number 198022 to 203742.

Notes

This dataset contains all runs from 2012 RunC. The list of validated runs, which must be applied to all analyses, can be found in

CMS list of validated runs Cert_190456-208686_8TeV_22Jan2013ReReco_Collisions12_JSON.txt

Characteristics

Dataset: **9828650** events **784** files **3.0 TB** in total

System Details

Global tag: FT_53_LV5_AN1

Recommended release for analysis: CMSSW_5_3_32

How were these data selected?

HLT trigger paths

The possible HLT trigger paths in this dataset are:

Export

JSON

How were these data selected?

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

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, photon, 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

Files

Filename	Size	Download	Preview
	3.8 GB		—
	2.4 GB		—
	4.0 GB		—
	3.6 GB		—
	3.8 GB		—

First | Previous | 1 | 2 | 3 | 4 | 5 | Next | Last

File Indexes

Filename	Size	Download	Preview
	59.6 kB		—
	33.8 kB		—

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.

The questions

Search and download

Search:

Possible to “tag” datasets

Search with those, e.g. 7TeV, type:PbPb, MC categories

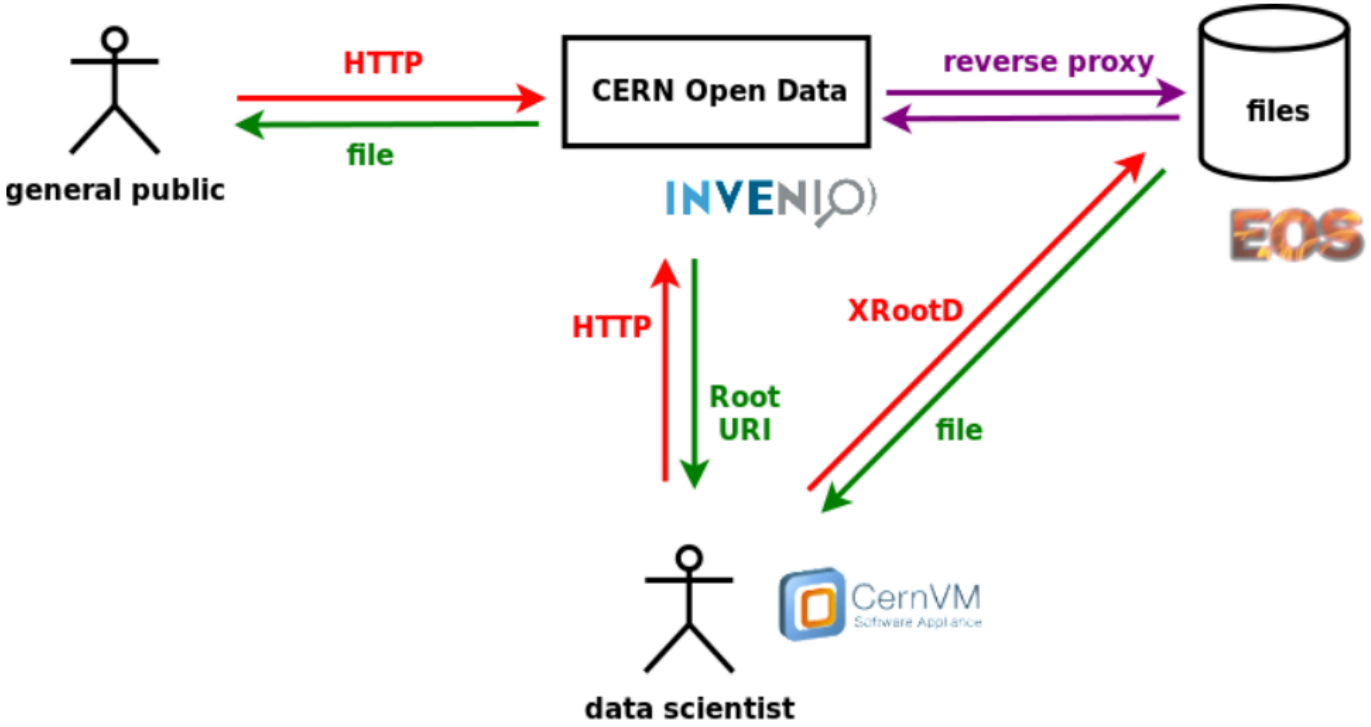
Downloading:

Each dataset is accessible by both

- HTTP streaming protocol
- XrootD streaming directly from EOS Public

CERN Open Data with EOS

Downloading:



http://opendata.cern.ch

https://github.com/cernopendata

DOIs

Essential for identifying a dataset and citing it!

But, no assignment to external datasets as we cannot guarantee immutability.

Once the DOI is minted, the dataset is “frozen” as is. BUT, we can:

version and have related datasets (is Parentof, isSupersetOf…)

Dataset size and format

Size:

Note: underlying storage is EOS

1TB releases not a problem

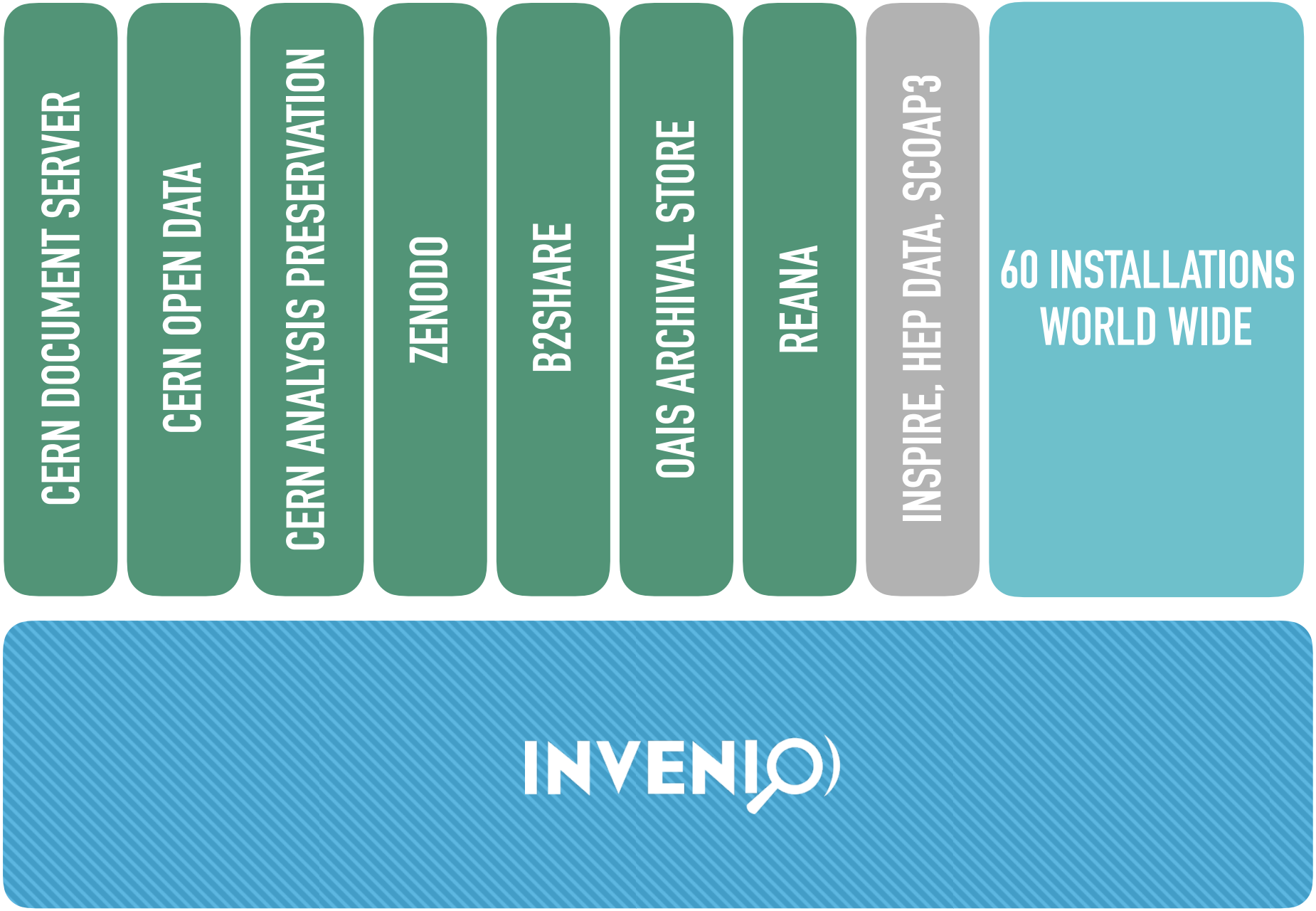
Format:

No limitations (but keep in mind the reuse)

One dataset identifier can contain several dataset formats

CERN Open Data in a nutshell

- particle physics oriented, customized data models
- tailored facets (energy, particles, …)
- LHC data so far, but also OPERA, perhaps LEP
- no self-depositing, data prepared with expert curation
- optimized for big releases
- only open content, no user accounts, no access controls
- both HTTP and XRootD streaming (direct EOSPUBLIC access)


CERN DOCUMENT SERVER
CERN OPEN DATA
CERN ANALYSIS PRESERVATION
ZENODO
B2SHARE
OAIS ARCHIVAL STORE
REANA
INSPIRE, HEP DATA, SCOAP3
60 INSTALLATIONS WORLD WIDE
INVENIO

http://opendata.cern.ch

https://github.com/cernopendata

Information structure


```
"publisher": "CERN Open Data Portal",
  "doi": "10.7483/OPENDATA.CMS.HJYH.UMWG",
  "license": {
     "attribution": "CC0"
  },
  "generator": {
     "global_tag": "START53_LV6",
     "names": [
        "madgraph"
     ]
  },
  "title": "/ZZJetsTo2L2Q_TuneZ2_7TeV-madgraph-tauola/Summer11LegDR-PU_S13_START53_LV6-v2/AODSIM",
  "control_number": "1308",
  "collision_information": {
     "energy": "7TeV",
     "type": "pp"
  },
  "relations": [
     {
        "type": "isChildOf",
        "title": "/ZZJetsTo2L2Q_TuneZ2_7TeV-madgraph-tauola/Summer11LegDR-PU_S13_START53_LV6-v2/
AODSIM"
     }
  ],
```

Information structure


```
{
   "files": [
      {
         "checksum": "sha1:c101dbd1bea6c4a0f4e0b02344693ff3c5d0272f",
         "uri": "root://eospublic.cern.ch//eos/opendata/cms/luminosity/2010/2010lumi.txt",
         "size": 63840
      },
      {
         "checksum": "sha1:47372391f417cec0e886fd912fccd37131530b1a",
         "uri": "root://eospublic.cern.ch//eos/opendata/cms/luminosity/2010/2010RunBlumi.txt",
         "size": 21429
      },
      {
         "checksum": "sha1:231d34287a8eb60320404a065a82a4caf8aa35f8",
         "uri": "root://eospublic.cern.ch//eos/opendata/cms/luminosity/2010/2010lumibyls.csv",
         "size": 5840441
      }
   ],
```