CERN (Open) Data Services

Tibor Šimko

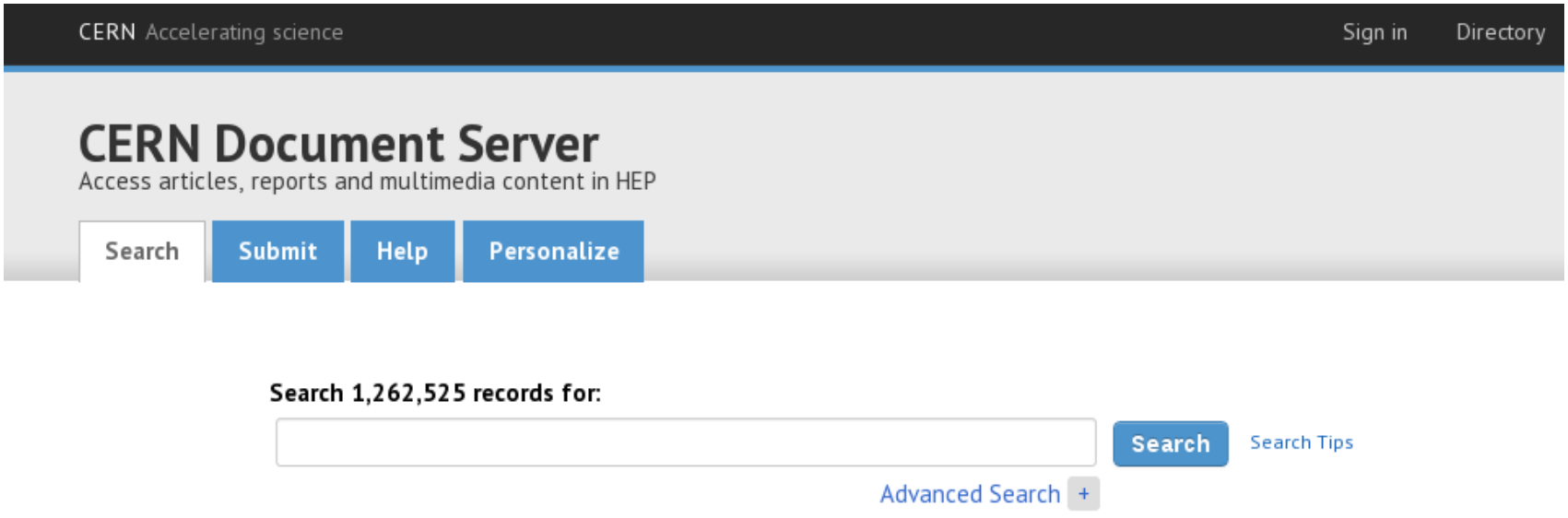
@tiborsimko

CERN-UNESCO School on Digital Libraries, Kumasi, Ghana · November 2016

“Small data”

CERN Document Server

cds.cern.ch


CERN Accelerating science
Sign in
Directory
CERN Document Server
Access articles, reports and multimedia content in HEP
Search
Submit
Help
Personalize
Search 1,262,525 records for:
Search
Search Tips
Advanced Search +
Articles & Preprints (1,086,540)
Published Articles (334,677) Preprints (671,836)
Theses (18,873) Reports (6,913)
CERN Notes (36,649)
Committee Documents (24,269)
Books & Proceedings (100,511)
Books (71,080) Proceedings (17,639)
Standards (11,639) Design Reports (171)
CERN Articles & Preprints (105,016)
CERN Published Articles (57,284) CERN Preprints (17,706) CERN Theses (4,690)
CERN Reports (1,180) Committee Documents (24,269)
CERN Series (18,729)
CERN Annual Reports (115) CERN Yellow Reports (1,150) CERN Theory (13,390)
Academic Training Lectures (706) Summer Student Lectures (951) General Talks (2,577)
CERN Departments (95,148)

“Related data files”

CERN Accelerating science — Sign in — Directory

CERN Document Server

Search | Submit | Help | Personalize

Home > Articles & Preprints > Published Articles > First observation of $\overline{B}^0 \to J/\psi K^+K^-$ and search for $\overline{B}^0 \to J/\psi\phi$ decays > Comments

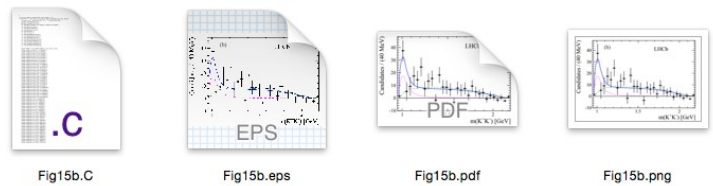
Information | Discussion (0) | Files

First observation of $\overline{B}^0 \to J/\psi K^+K^-$ and search for $\overline{B}^0 \to J/\psi\phi$ decays - Aaij, R *et al* - arXiv:1308.5916

Main file(s):

prd.88.e072005

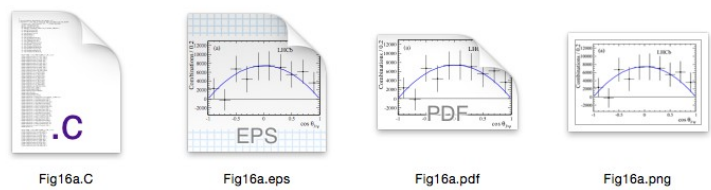
version 1 prd.88.e072005.pdf [1.32 MB] *07 Nov 2013, 15:17* *APS Open Acc*



Additional file(s):

Related data file(s)

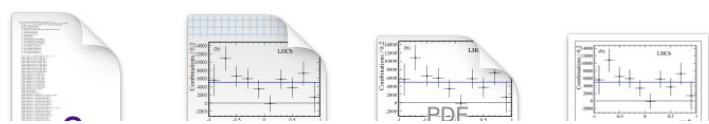
version 1 Related data file(s).zip [10.36 MB] *02 Sep 2013, 16:41*



arXiv file(s):

arXiv:1308.5916

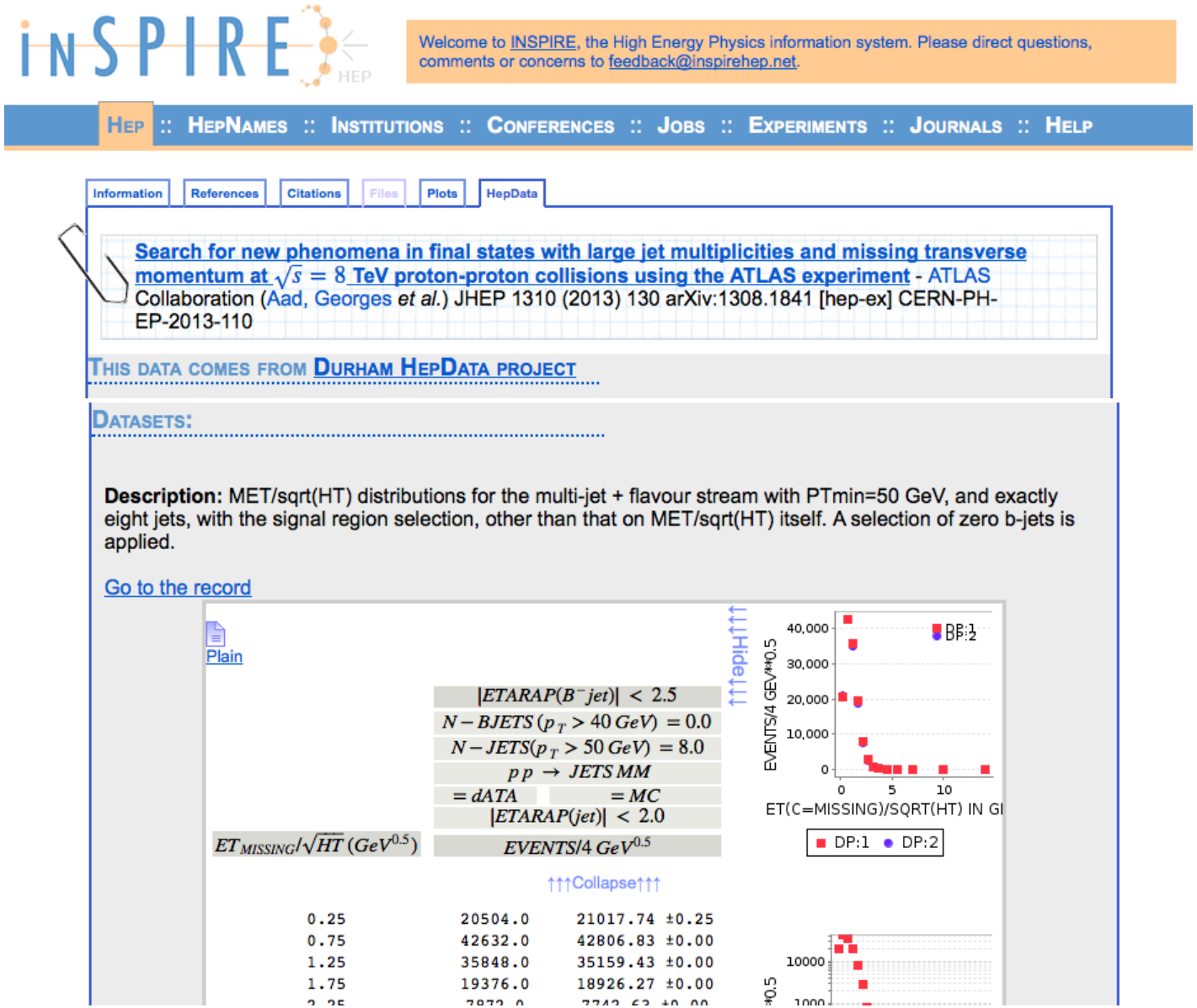
version 4 (see previous) arXiv:1308.5916.pdf [4.36 MB] *26 Oct 2013, 04:15*



INSPIRE

inspirehep.net

“Data behind plots”



HEPDATA

“Interactive data behind plots”

HEPDATA Search HEP Data Search Submit Sandbox Help Your Account

Browse all | Cohen, Martin et al.

Ultra- and Hyper-compact HII regions at 20 GHz

Cohen, Martin, Murphy, Tara, Ekers, Ronald D., Green, Anne J., Wark, Robin, Moss, Vanessa

Abstract
We present radio and infrared observations of 4 hyper-compact HII regions and 4 ultra-compact HII regions in the southern Galactic plane. These objects were selected from a blind survey for UCHII regions using data from two new radio surveys of the southern sky/ the Australia Telescope 20 GHz survey (AT20G) and the 2nd epoch Molonglo Galactic Plane Survey (MGPS-2) at 843 MHz. To our knowledge, this is the first blind radio survey for hyper- and ultra-compact HII regions. We have followed up these sources with the Australia Telescope Compact Array to obtain H70-alpha recombination line measurements, higher resolution images at 20 GHz and flux density measurements at 30, 40 and 95 GHz. From
Data Abstract
CERN-LHC. Measurements of the cross section for ZZ production using the 4l and 2l2nu decay channels in proton-proton collisions at a centre-of-mass energy of 7 TeV with 4.6 fb^-1 of data collected in 2011. The final states used are 4 electrons, 4 muons, 2 electrons and 2 muons, 2 electrons and missing transverse momentum, and 2 muons and missing transverse momentum (MET). The cross section values reported in the tables should be multiplied by a factor of 1.0141 to take into account the updated value of the integrated luminosity for the ATLAS 2011 data taking period. The uncertainty on the global normalisation ("Lumi") remains at 1.8%. See Eur.Phys.J. C73 (2013) 2518 for more details. The 4l channel fiducial region is

DOI 10.1111/j.1365-2966.2010.16589.x | View in Inspire

combined systematic uncertainty excluding luminosity, the second is the...
passed review

Table 3
Normalized ZZ fiducial cross section (multiplied by 10^6 for readability) in bins of the leading reconstructed dilepton pT for the...
passed review

Table 4
Normalized ZZ fiducial cross section (multiplied by 10^6 for readability) in bins of the dilepton pT for the 2l2nu channel....
passed review

Table 5
Normalized ZZ fiducial cross section (multiplied by 10^6 for readability) in bins of deltaPhi between the two leptons of the...
passed review

Table 8
Normalized ZZ fiducial cross section (multiplied by 10^6 for readability) in bins of the transverse mass of the ZZ system...

Normalized ZZ fiducial cross section (multiplied by 10^6 for readability) in bins of deltaPhi between the two leptons of the leading dileptons for the 4l channel. The first systematic uncertainty is detector systematics, the second is background systematic uncertainties. UPDATE (30 APR 2014): extra significant digit added for first bin

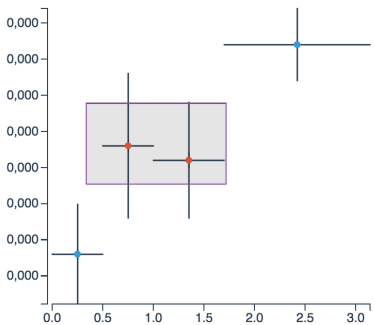
energies 7000 | observables DSIG/DPHI | reactions P P --> Z0 Z0 X

Data

Download Data

RE	P P --> Z0 < LEPTON+ LEPTON- > Z0 < LEPTON+ LEPTON- > X
SQRT(S)	7000
Leading dilepton DELTA(PHI(LEPTON+, LEPTON\-)) GEV	10**6 * 1/SIG(fiducial) * D(SIG(fiducial))/DDELTA(PHI(LEPTON+,LEPTON-))
0.5 - 1	280000 100000, 100000 stat 9000, 9000 sys,detector 400, 400 sys,background
1 - 1.7	260000 80000, 80000 stat 10000, 10000 sys,detector 300, 300 sys,background

Visualize



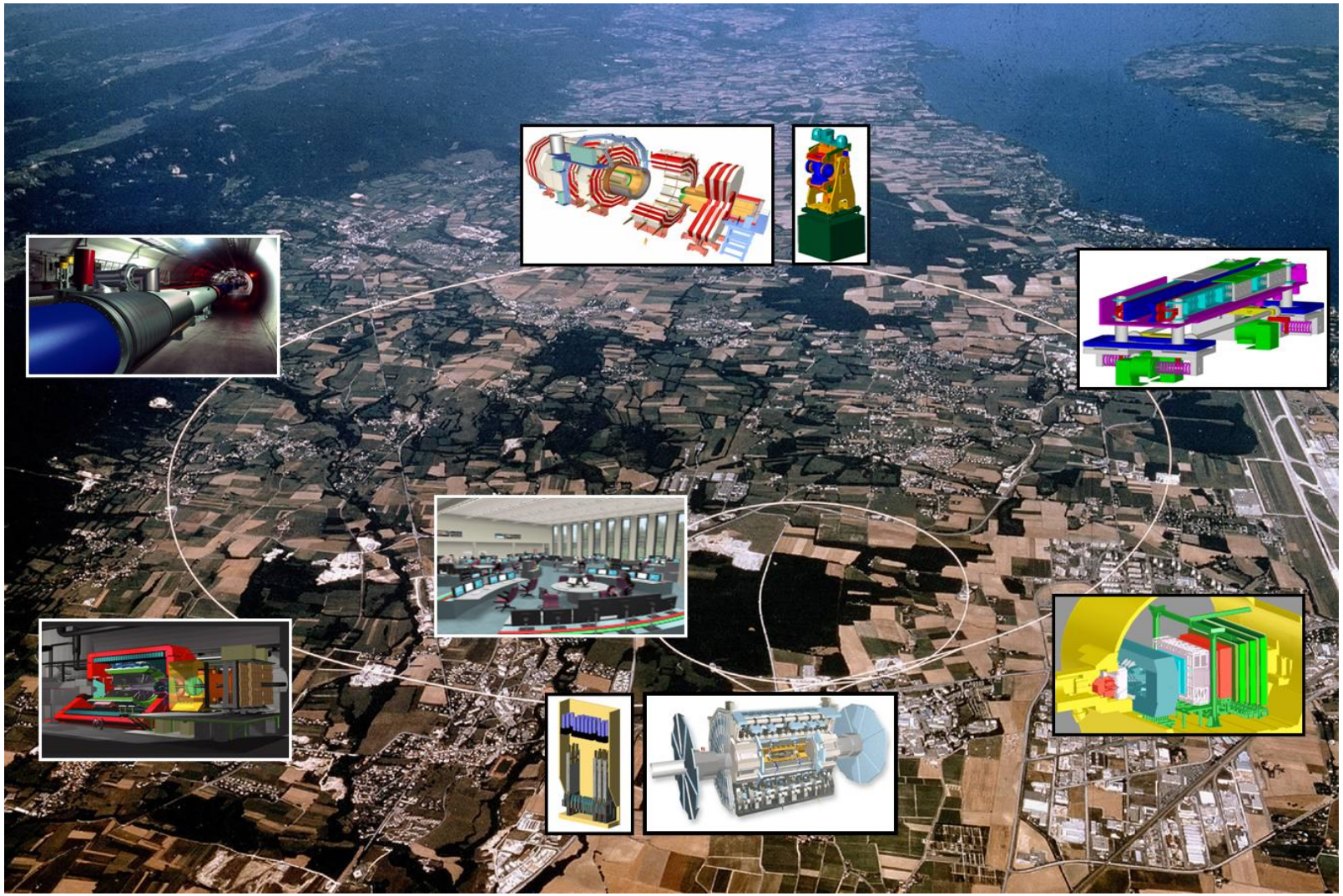
Deselect variables or hide different error bars by clicking on them.

Variables
10**6 * 1/SIG(fiducial) * D(SIG(fiducial))/DDELTA(PHI(LEPTON+,LEPTON-))
stat error
sys,detector error

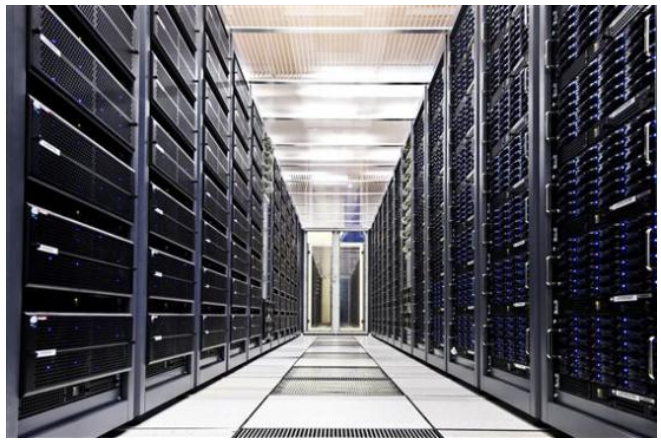
Copyright ~1975-Present, HEP Data | Powered by Invenio
About Team Contact

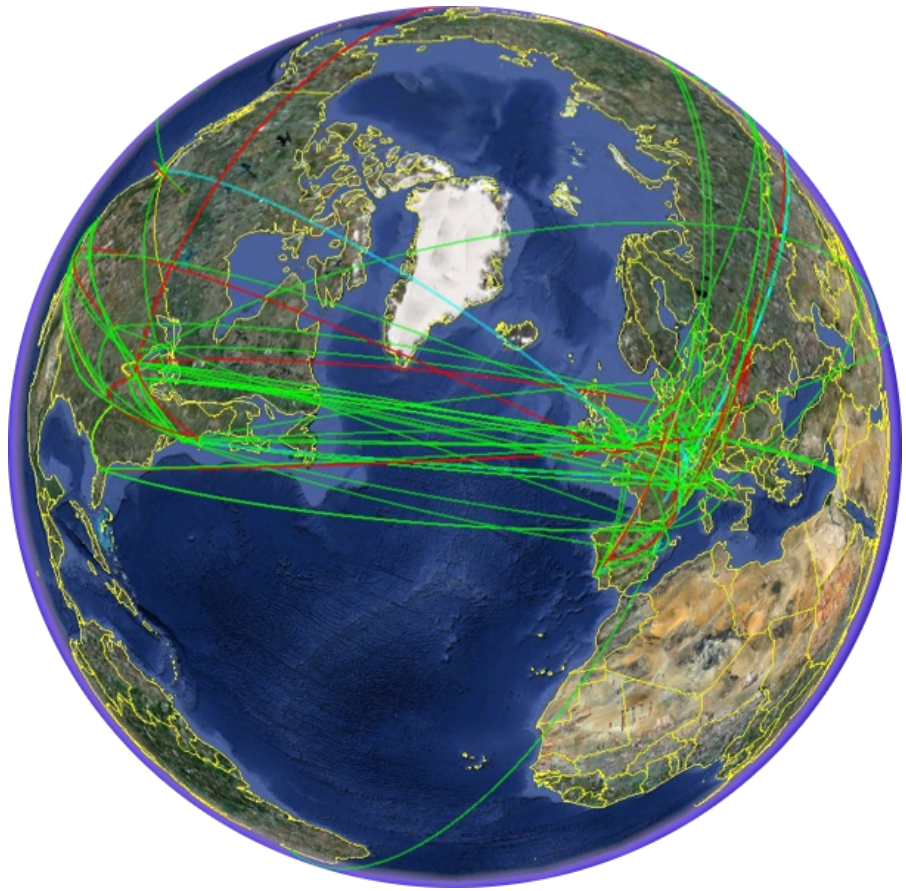
“Big data”

CERN LHC Experiments



Large Scale Solutions





Primary site: 100k cores (10k nodes), 100k disks (50 PB), 21k NIC

Grid: 13 Tier-1 sites, 155 Tier-2 sites, 10 Gbps optical fibre

LHC Data Pyramid

~GB/analysis

~TB/analysis

~PB/year

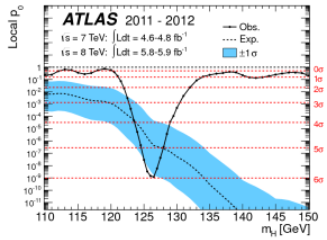
~GB/sec

1. Provision of additional documentation for the published results

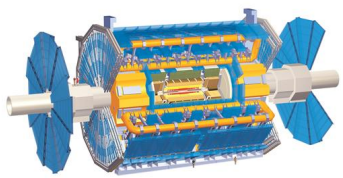
2. Simplified data formats for analysis in outreach and training exercises

3. Reconstructed data and simulations as well as the analysis level software to allow a full scientific analysis

4. Basic raw level data (if not yet covered as level 3 data) and their associated software which allows access to the full potential of the experimental data

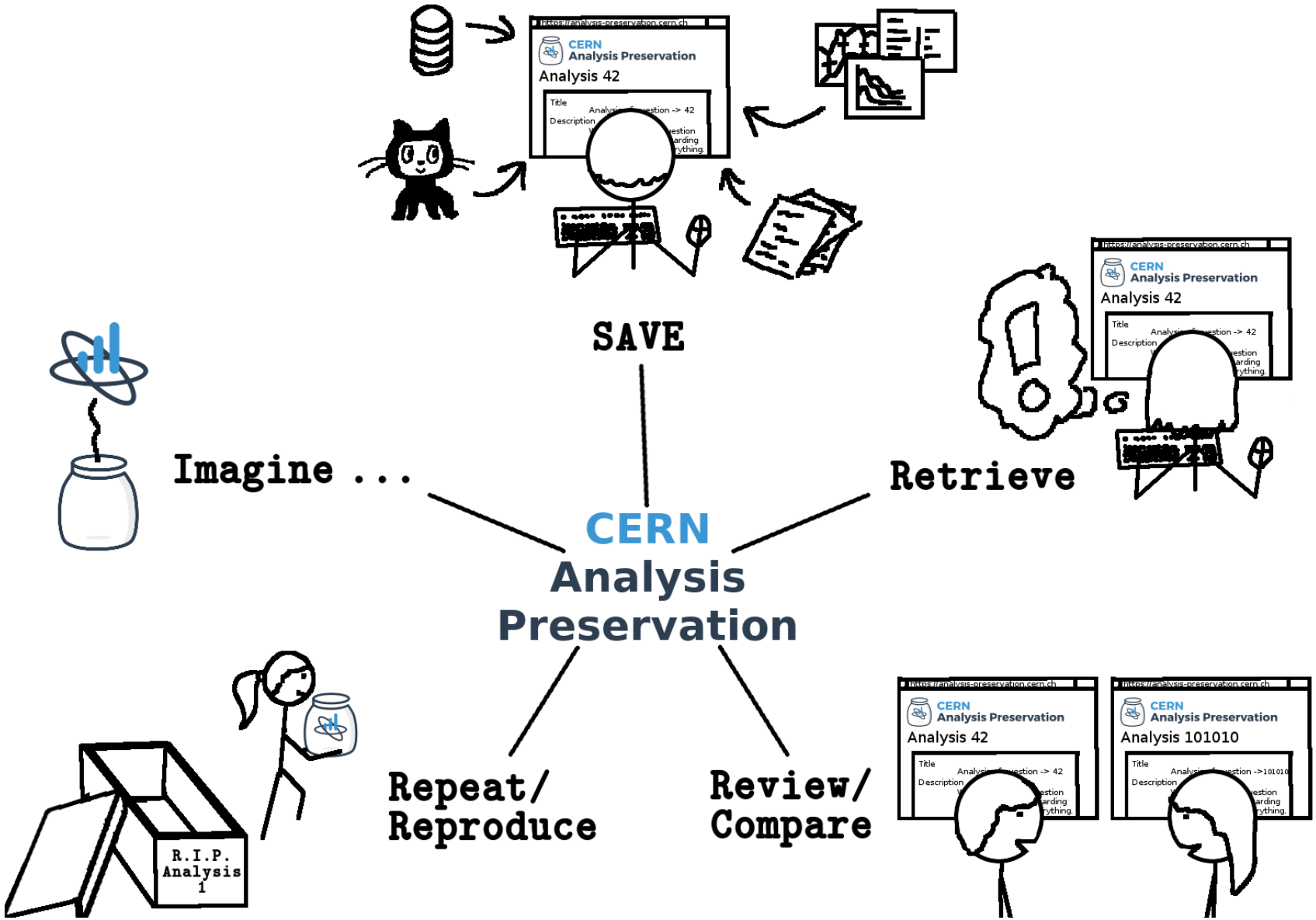


analysis

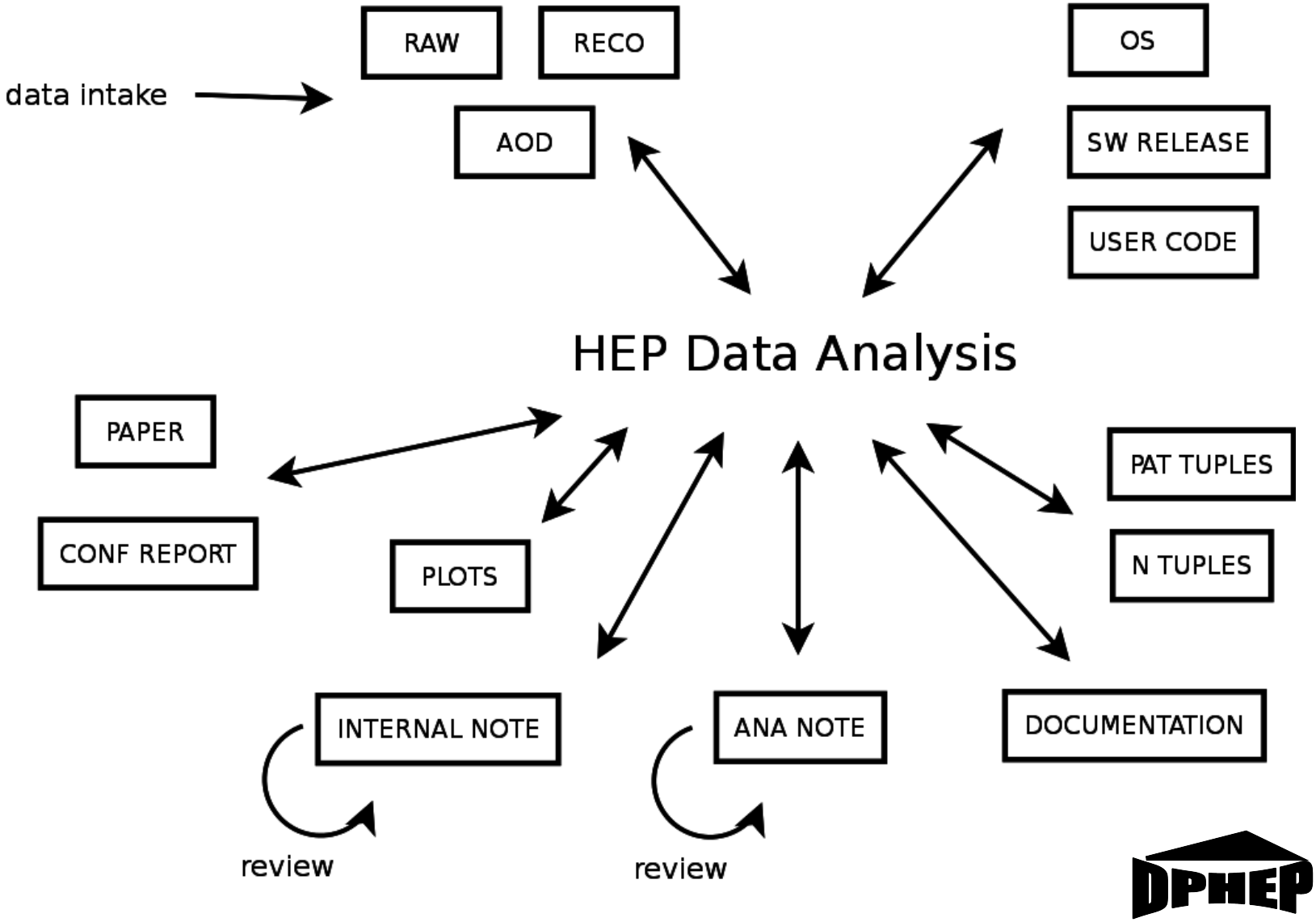


CERN Analysis Preservation

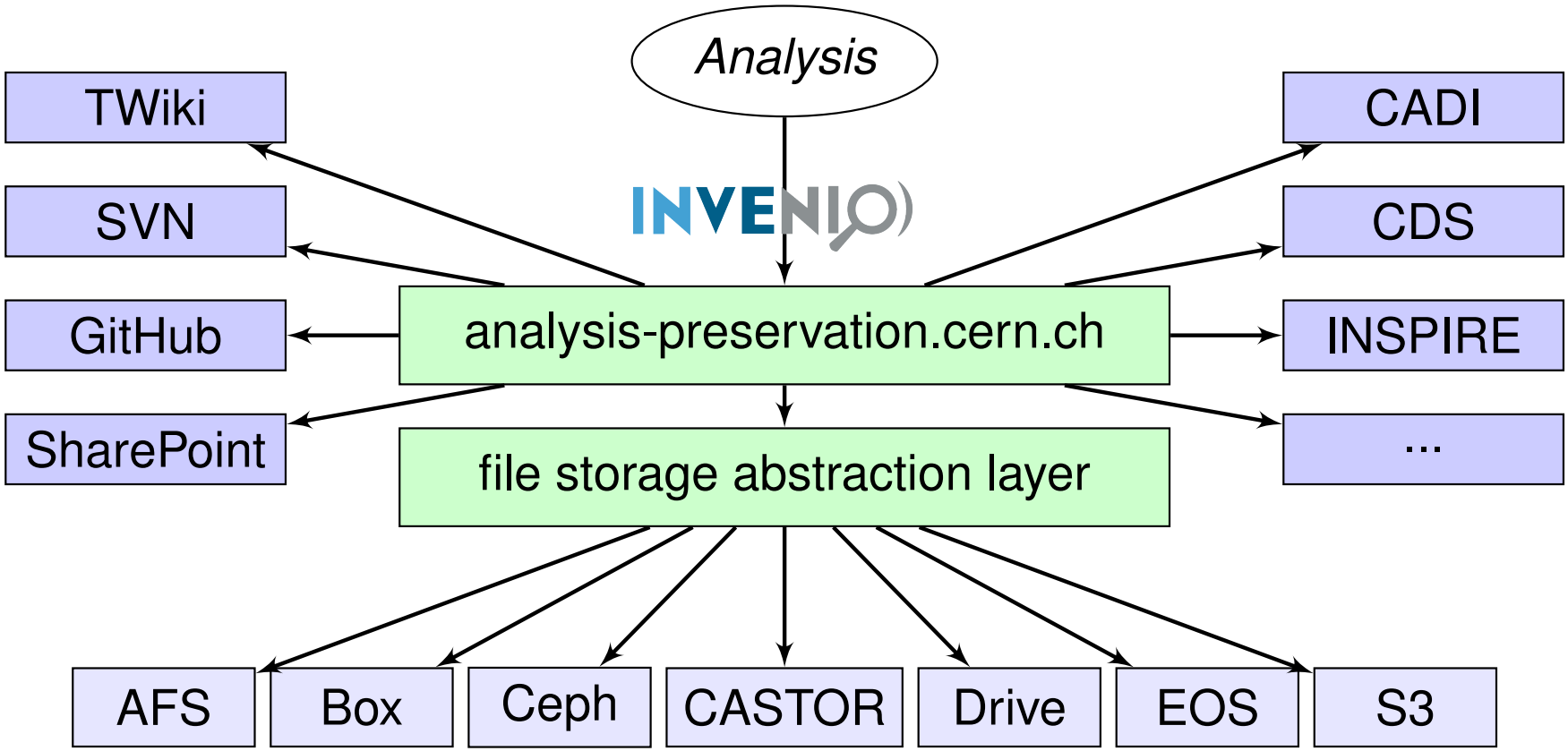
User stories



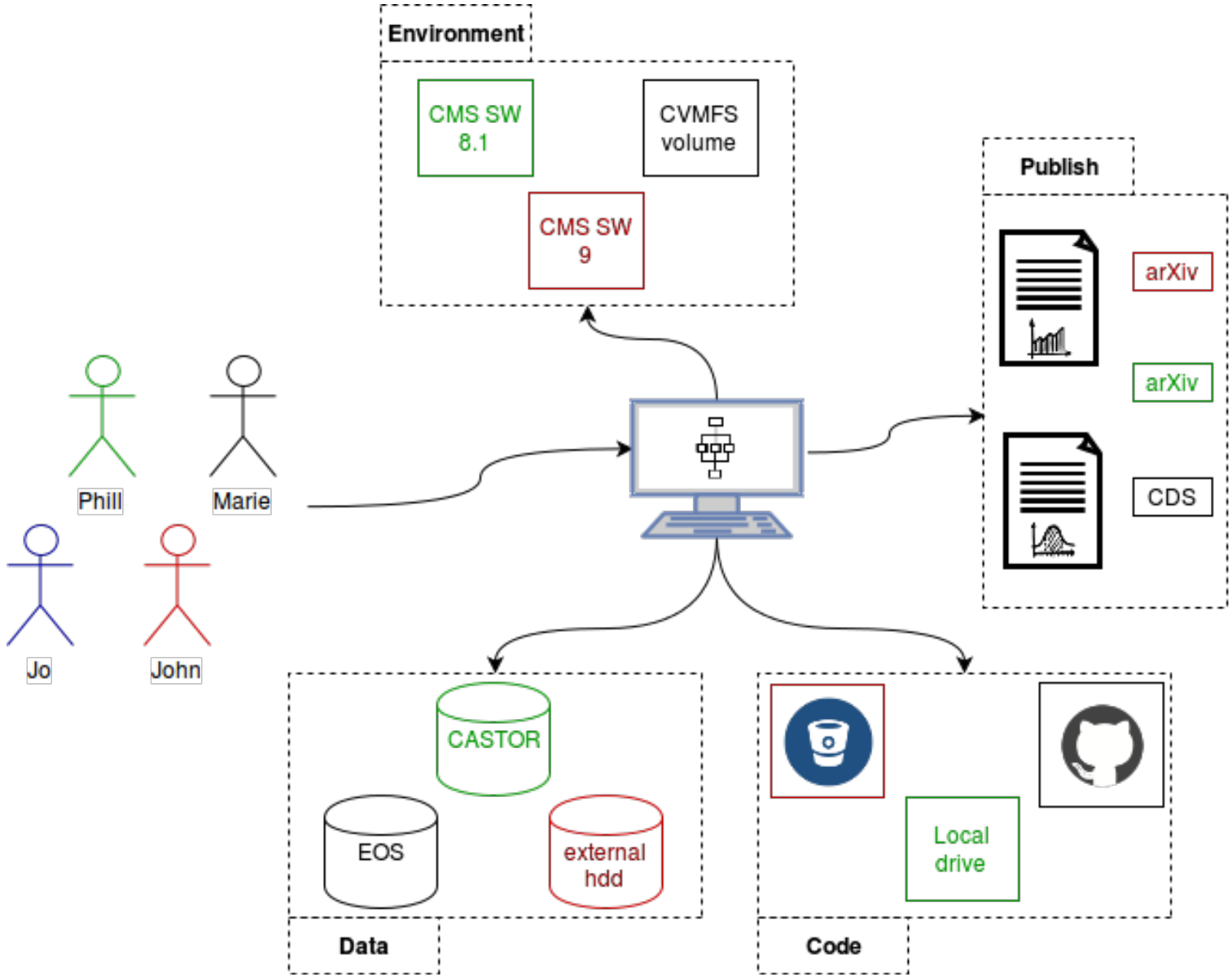
Preserve an analysis?



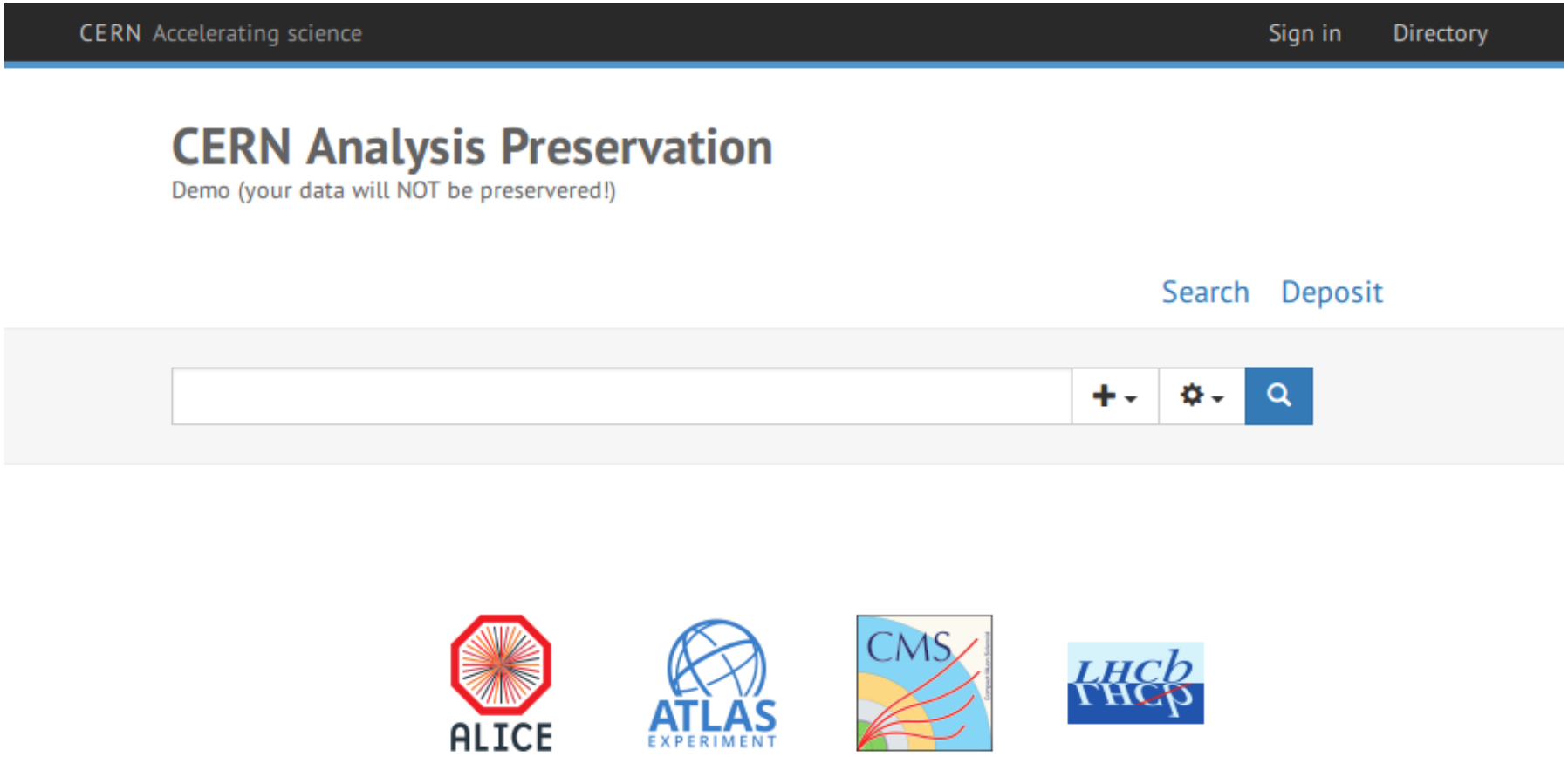
System architecture



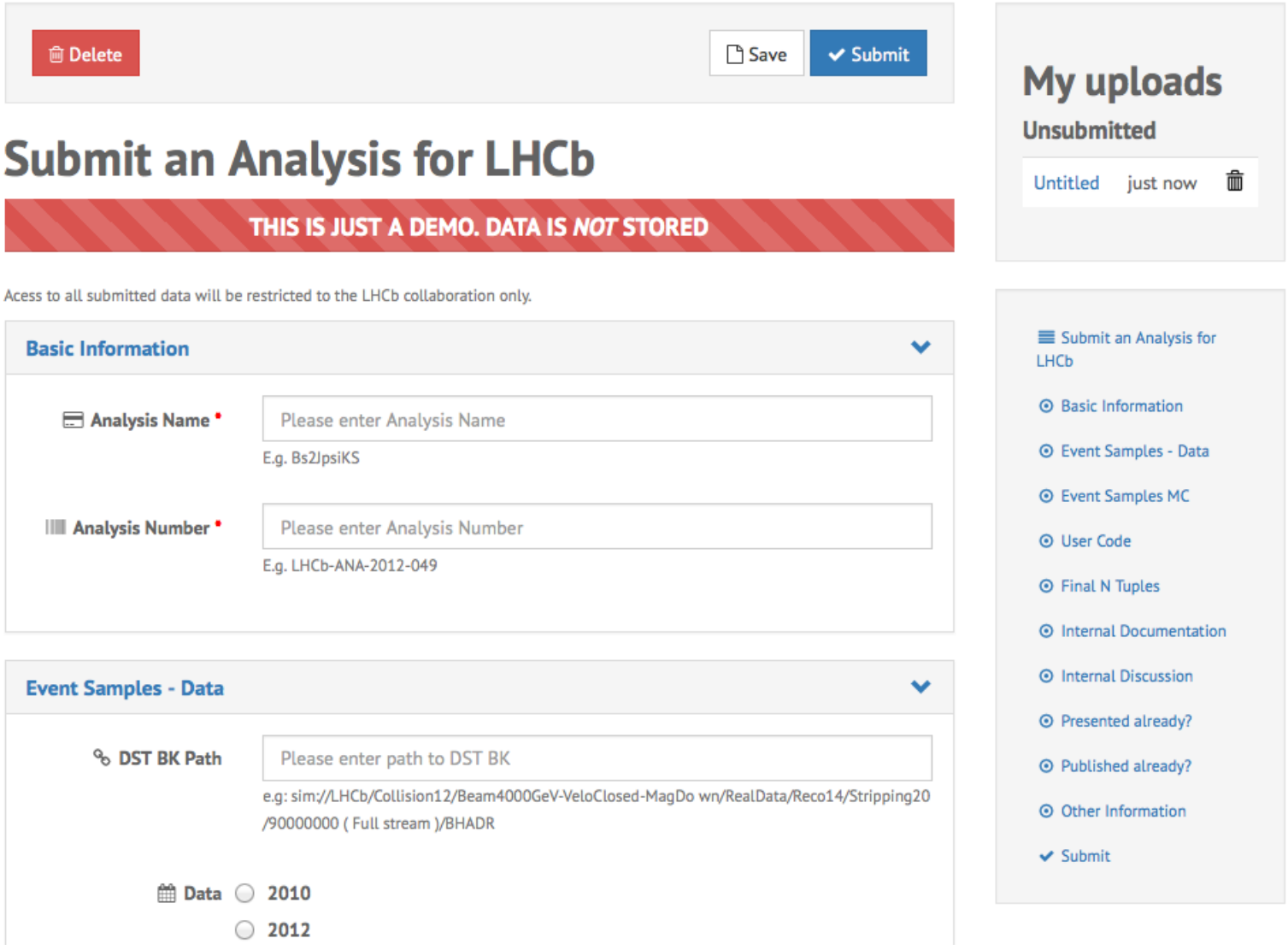
Challenges


Environment
CMS SW 8.1
CVMFS volume
CMS SW 9
Publish
arXiv
arXiv
CDS
Phill
Marie
Jo
John
CASTOR
EOS
external hdd
Local drive
Data
Code

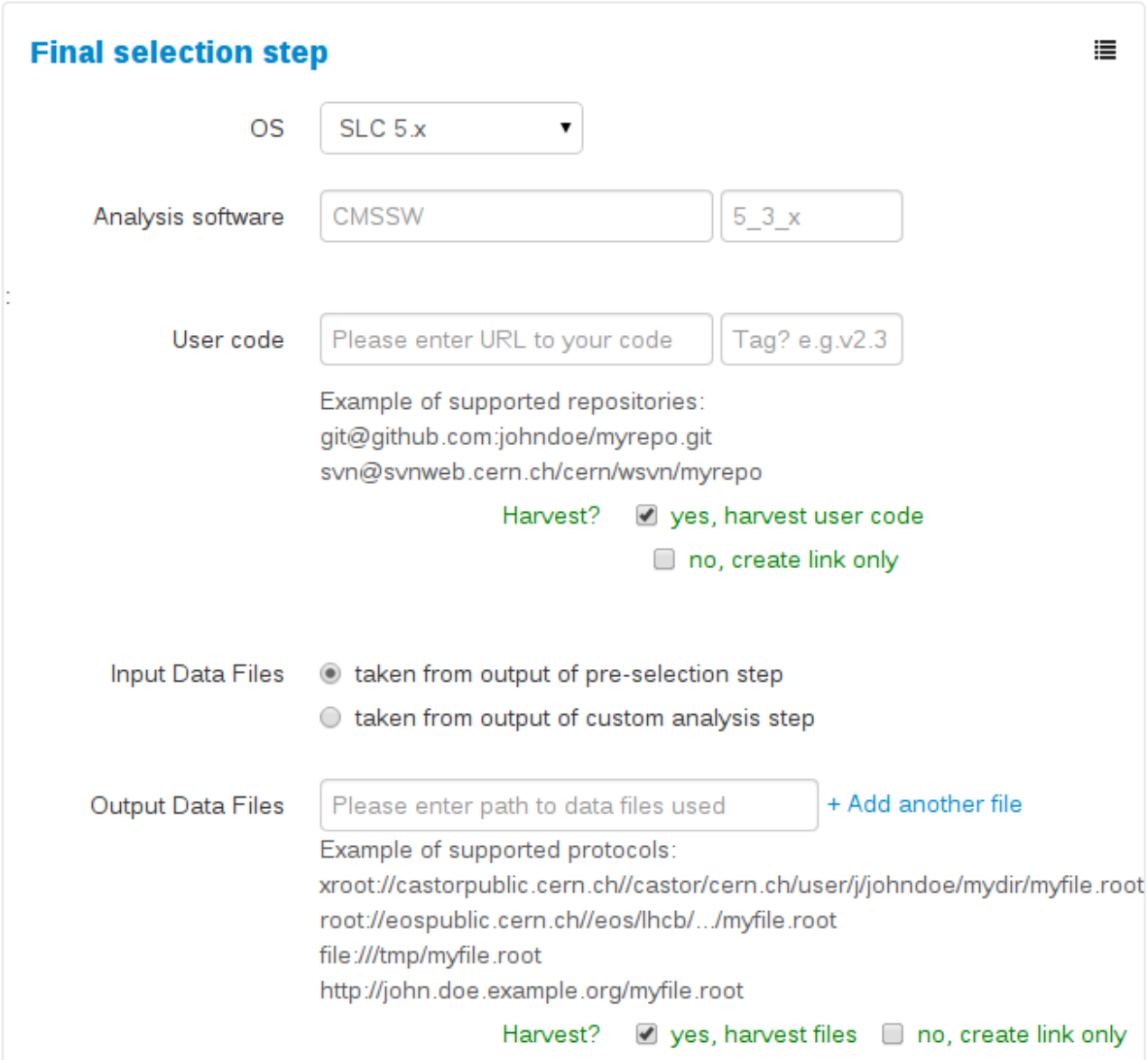
Pilot project



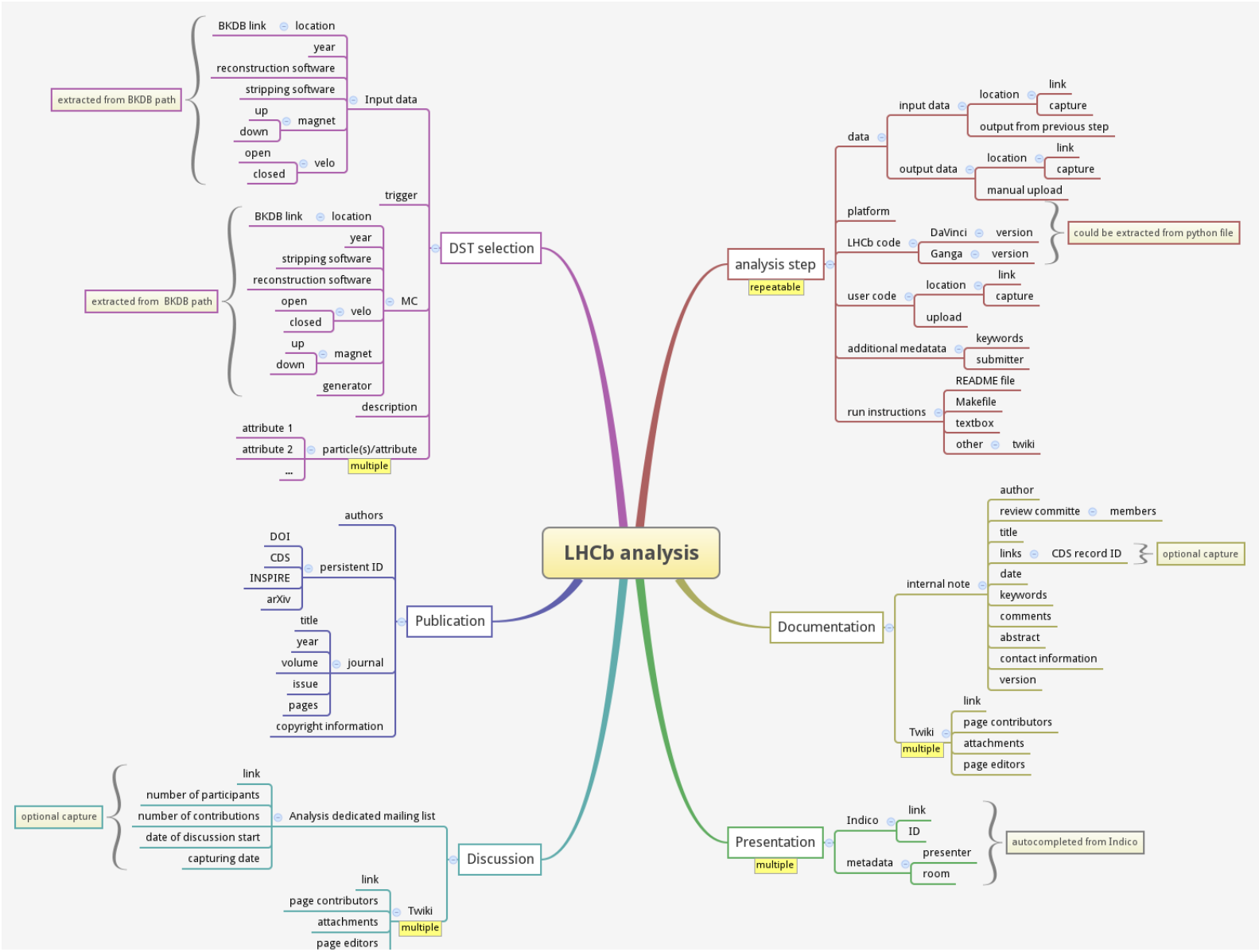
Example: LHCb analysis



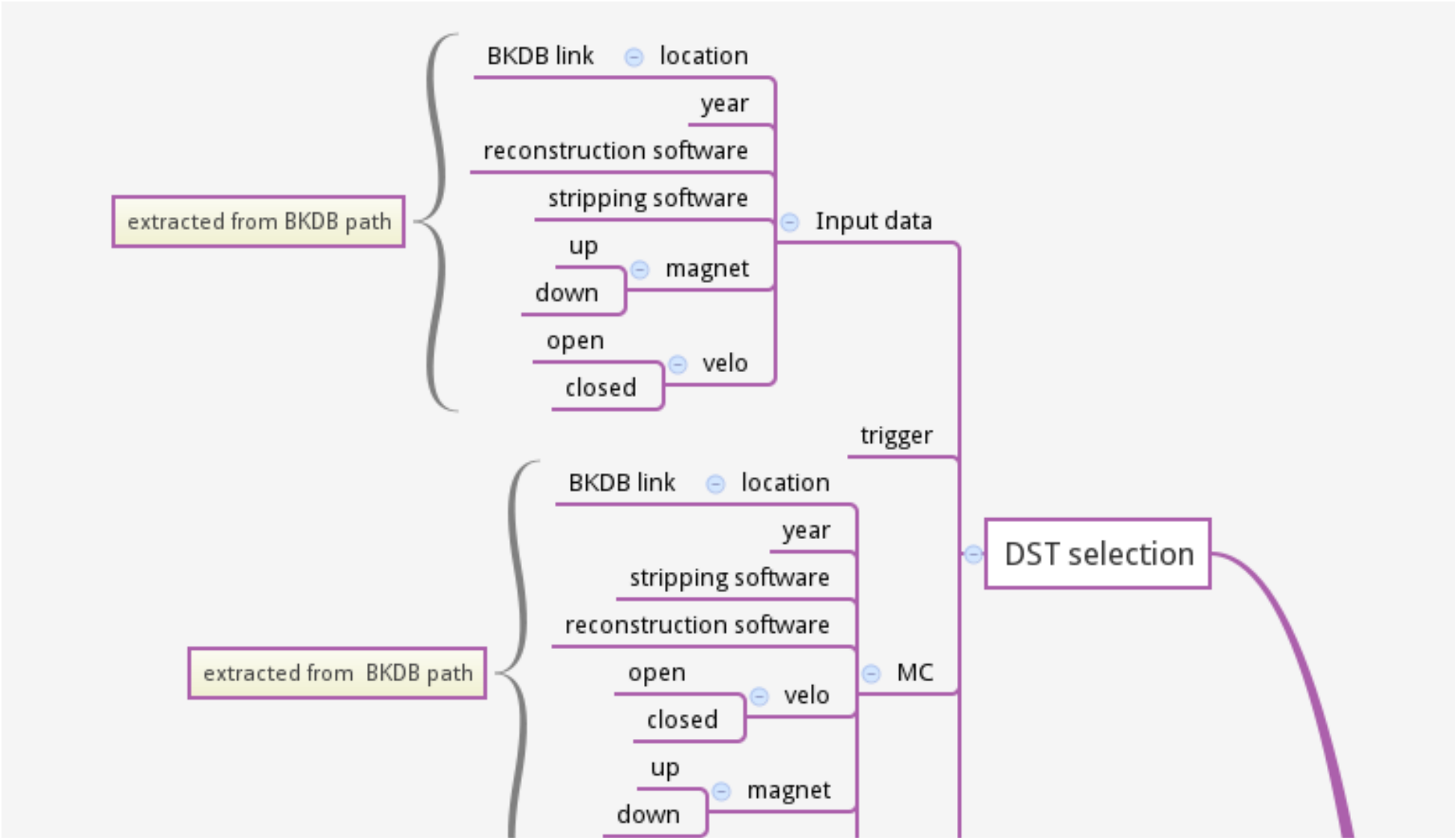
Knowledge capture



Knowledge modelling



Knowledge modelling

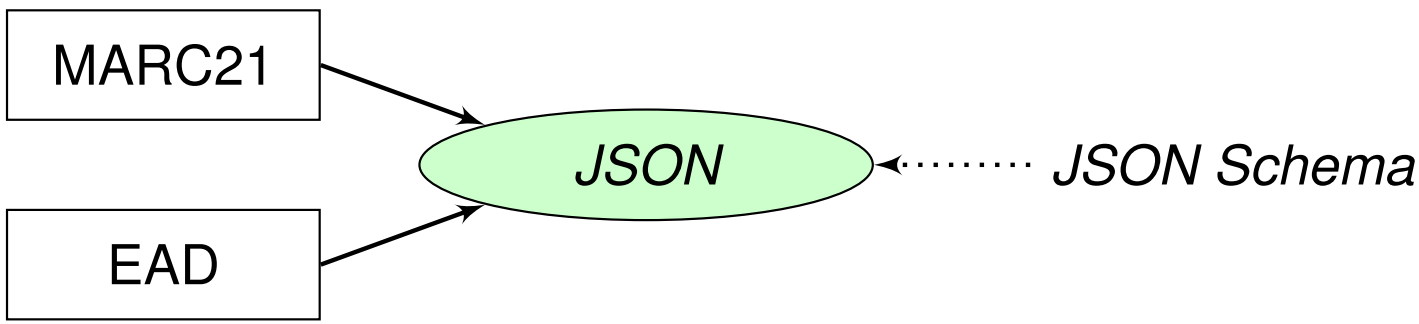


Knowledge representation

- prototype: extended MARC21
  - “technical” metadata: beyond bytes

    e.g. `256 ‘‘computer file characteristics’’`
    ```
    $a characteristics   $e events      $t text
    $b bytes             $f files        ...
    ```
  - “knowledge” metadata: semantics

    e.g. `505 ‘‘formatted contents note’’ CSV column information`
    ```
    $t title      $g miscellaneous
    ```
- internal format: JSON



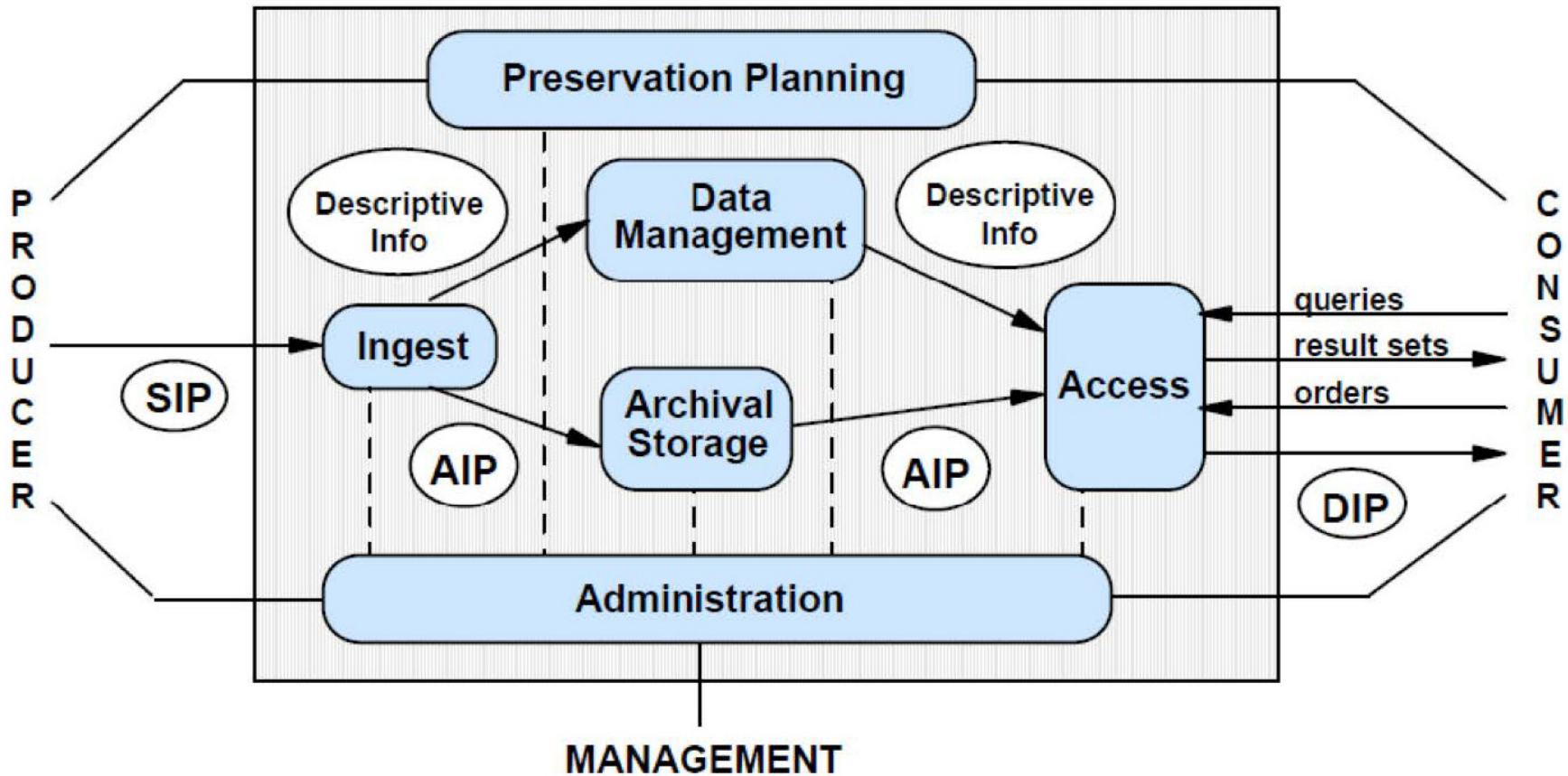
json-schema.org

```
{
  "title": "Example Schema",
  "type": "object",
  "properties": {
    "firstName": {
      "type": "string"
    },
    "lastName": {
      "type": "string"
    },
    "age": {
      "description": "Age in years",
      "type": "integer",
      "minimum": 0
    }
  },
  "required": ["firstName", "lastName"]
}
```

Datasets in JSON

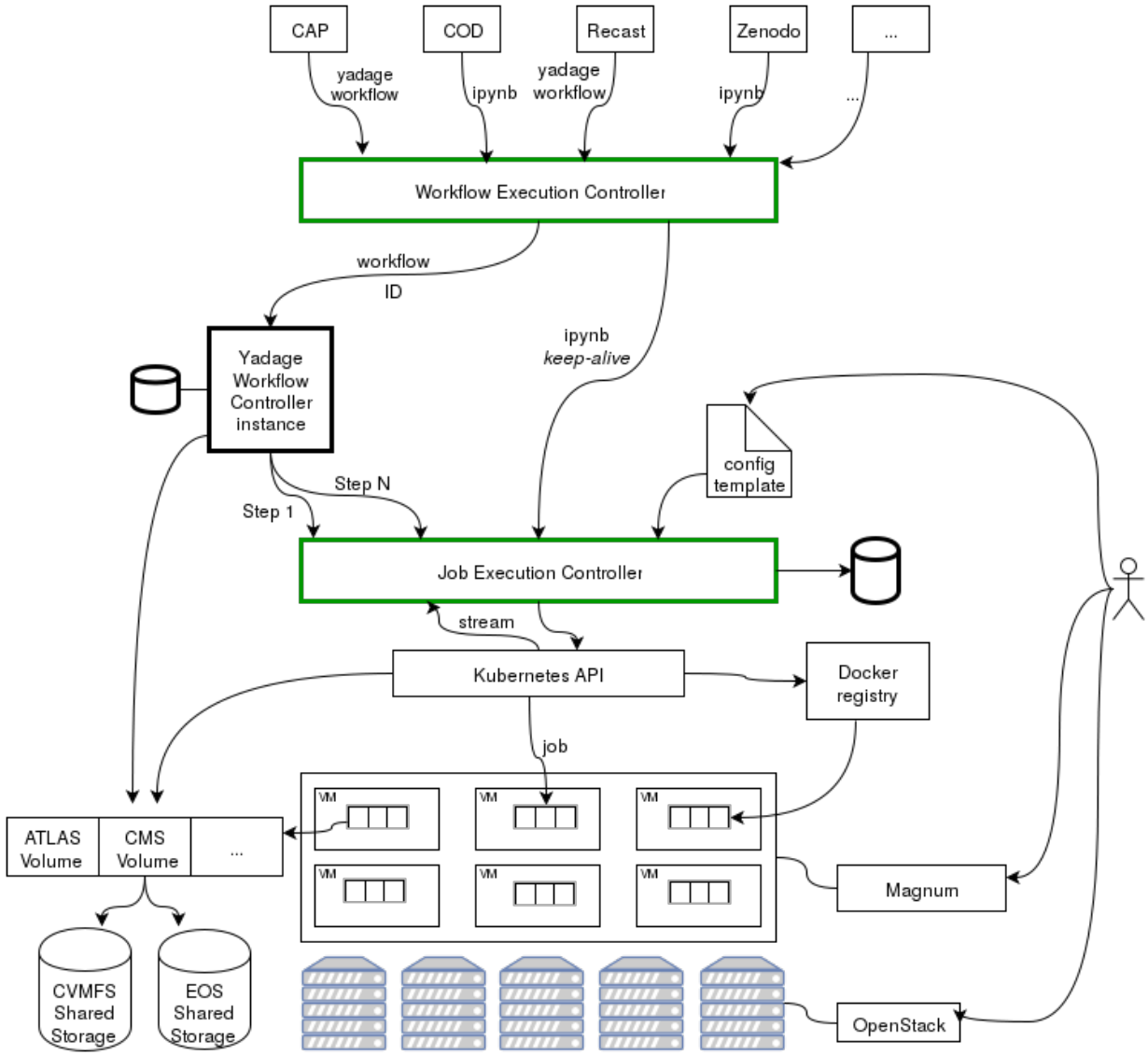
```
"primary_dataset": [
  {
    "@type": "dcat:Dataset",
    "title": "/Mu/Run2010B-Apr21ReReco-v1/AOD",
    "description": "Mu primary dataset in AOD format from
    "licence": "CC0 waiver",
    "persistent_identifiers": [
      {
        "identifier": "10.7483/OPENDATA.CMS.B8MR.C4A2",
        "scheme": "DOI"
      }
    ],
    "issued": "2011-04-26 11:32:43",
    "modified": "2011-05-02 21:22:30",
```

Open Archival Info System



SIP = Submission Information Package · AIP = Archival Information Package · DIP = Dissemination Information Package

Re-run preserved analysis?



CERN Open Data

Open data policies

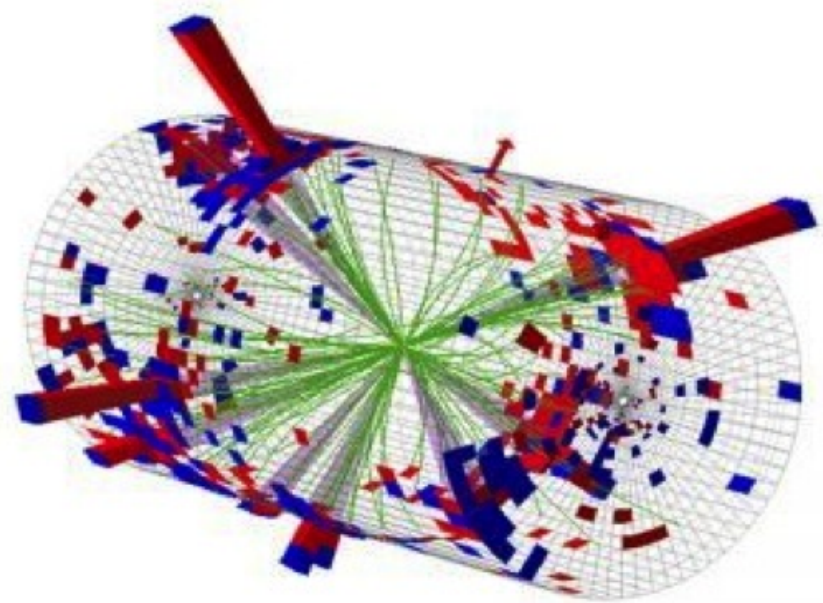
Data policies

- restricted $\rightarrow$ embargo period $\rightarrow$ open

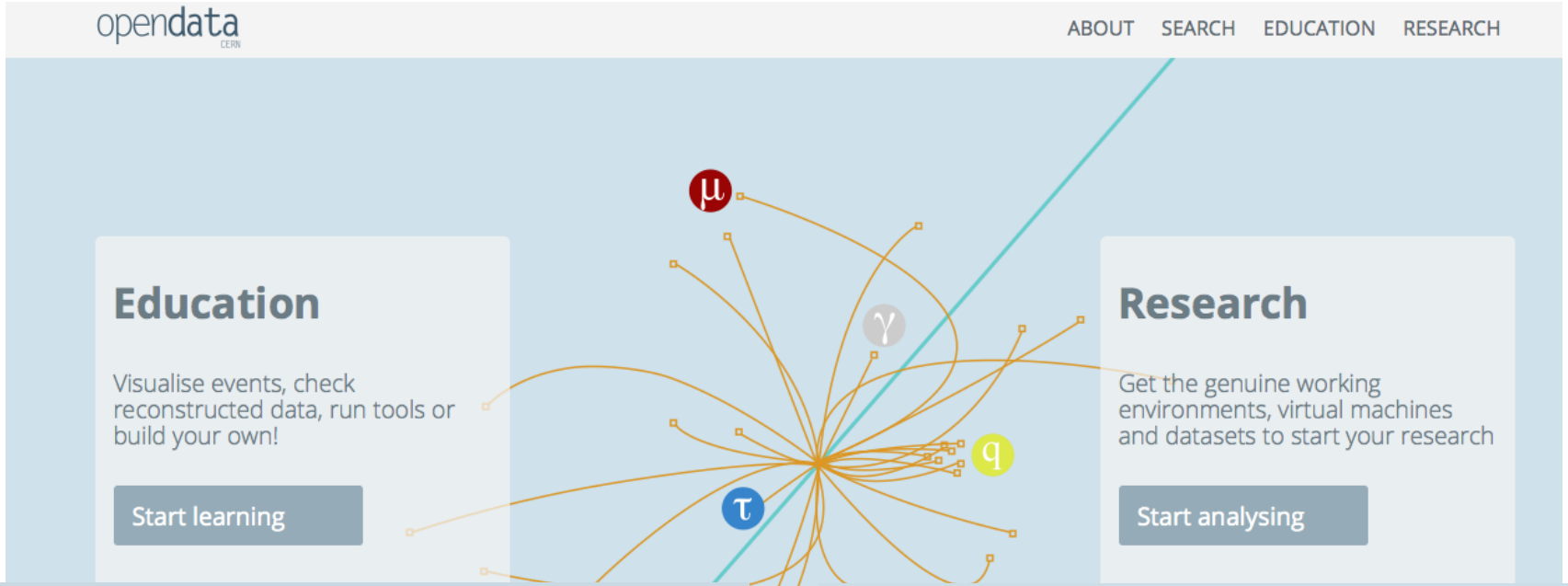
  "[...] Data with high abstraction, such as AOD, will be conditionally made publicly available after an embargo period of 5 years after publication for 10% of the data and 10 years for 100% of the data [...]" —ALICE Data Policy

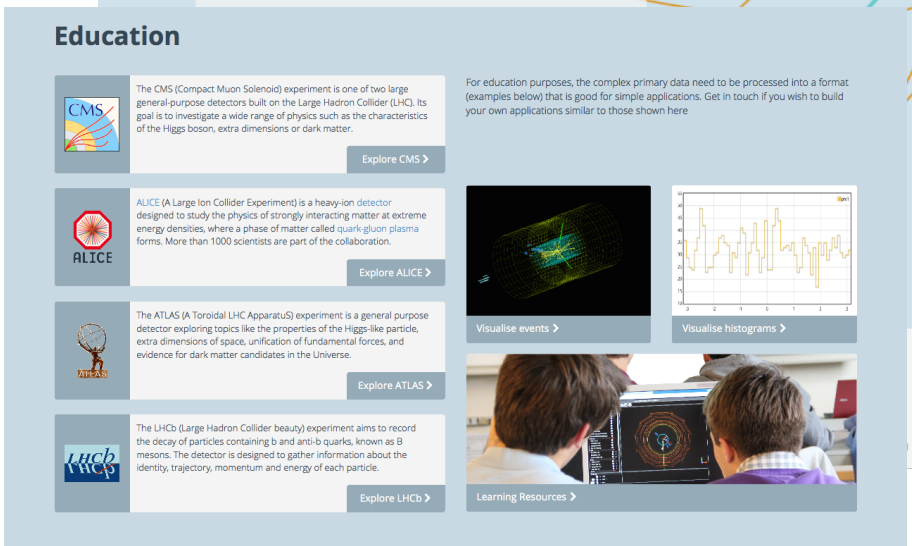
Challenges

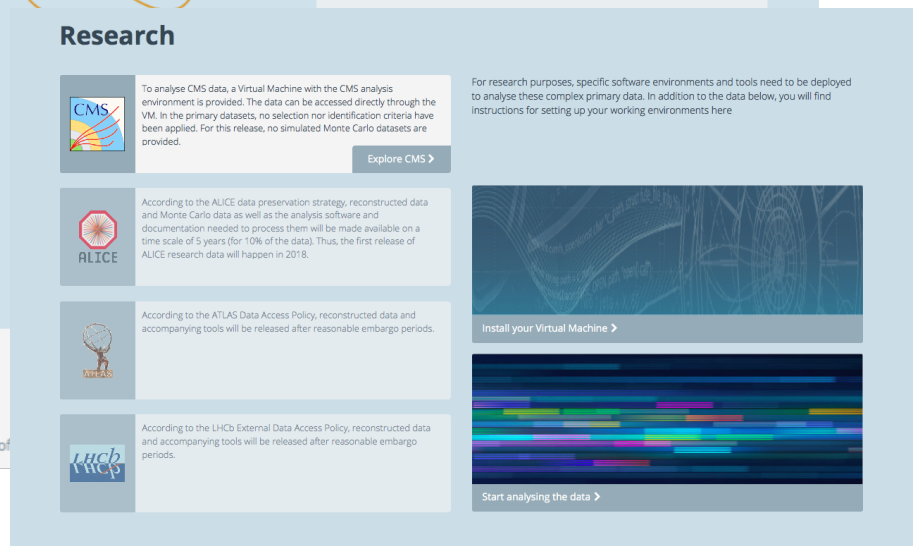
- audience
  - general public
  - high-school students
  - citizen scientists
  - data miners
- computing
  - exploring in the browser
  - using specialised VMs



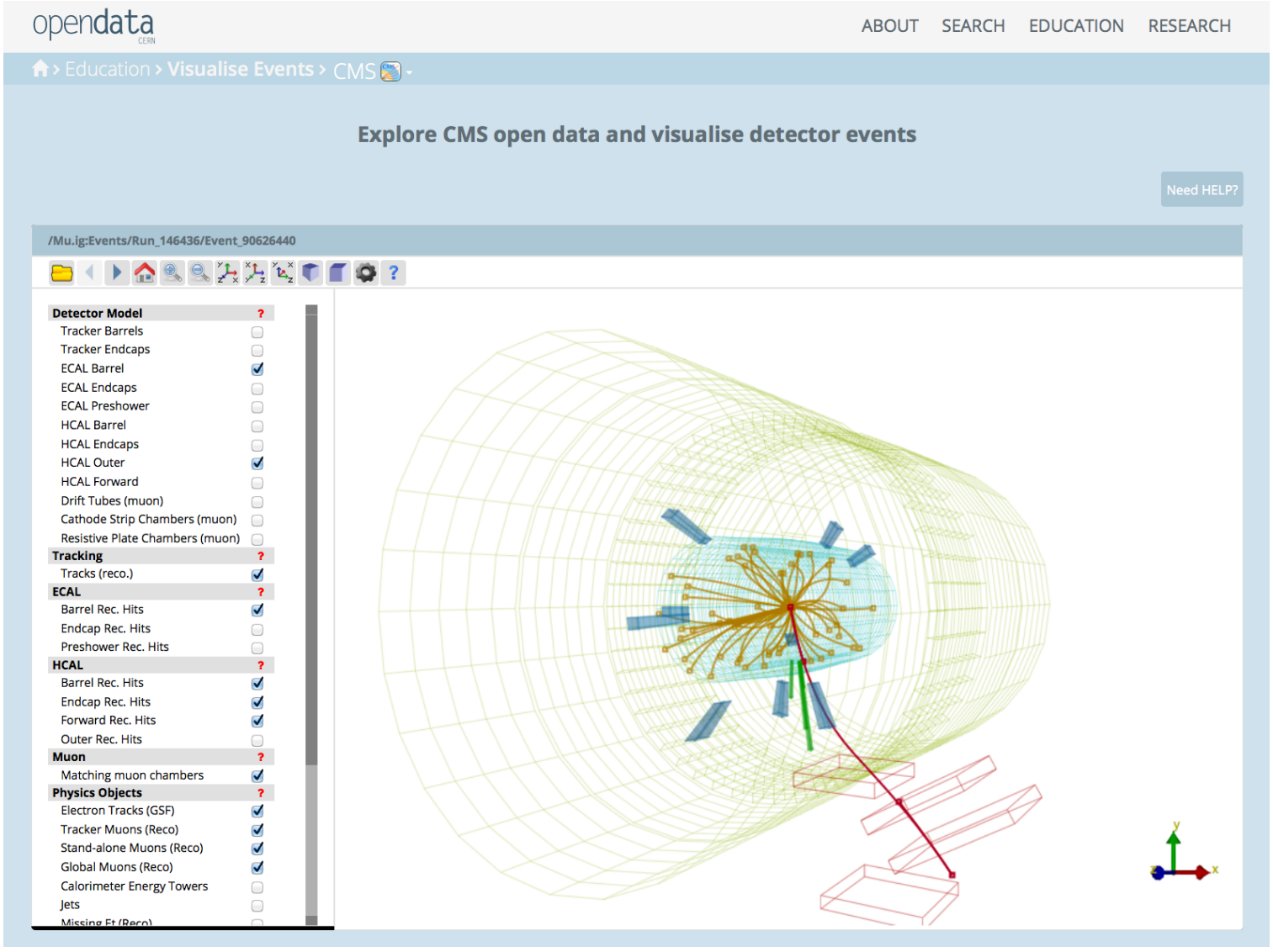
opendata.cern.ch



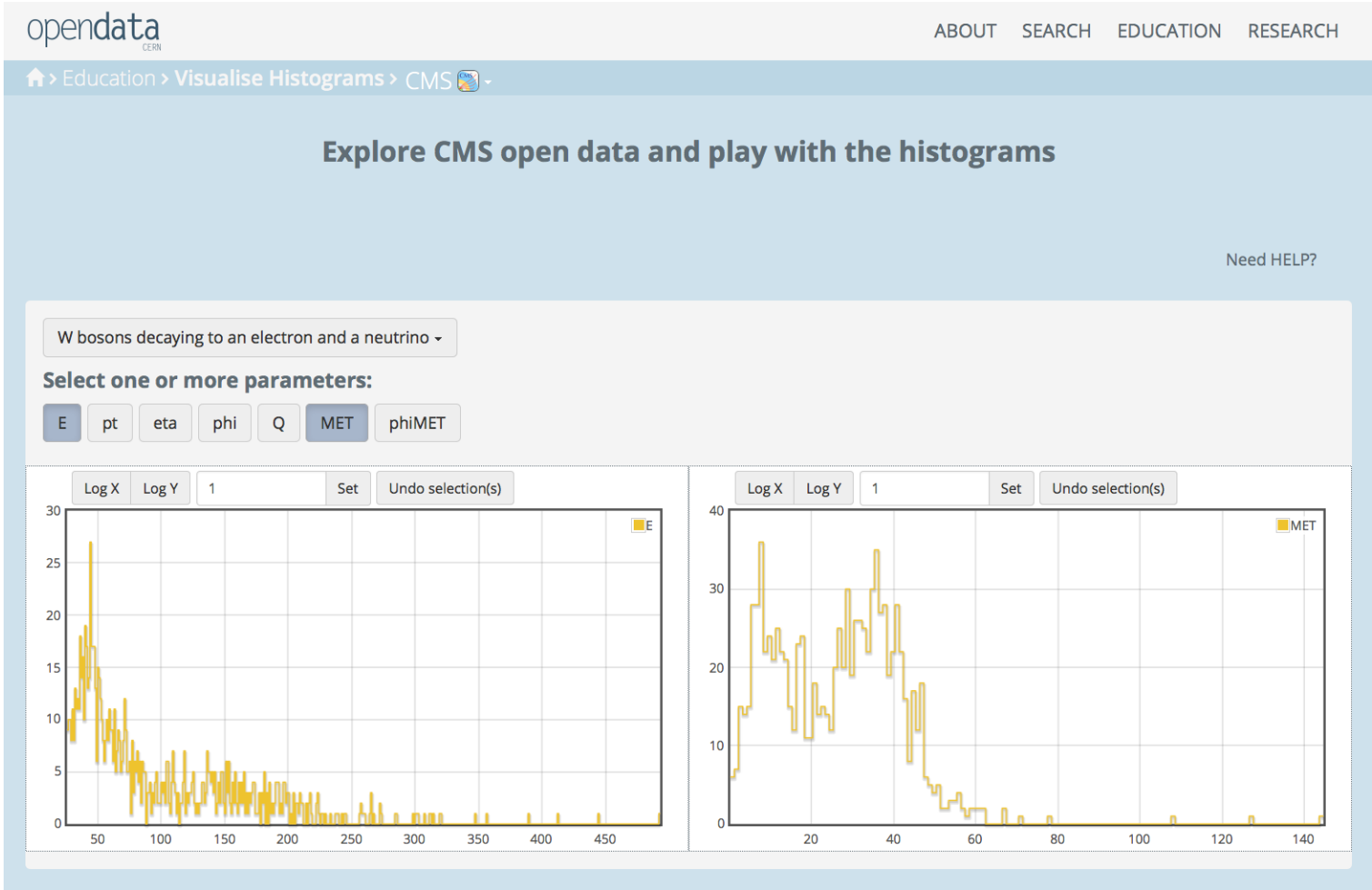




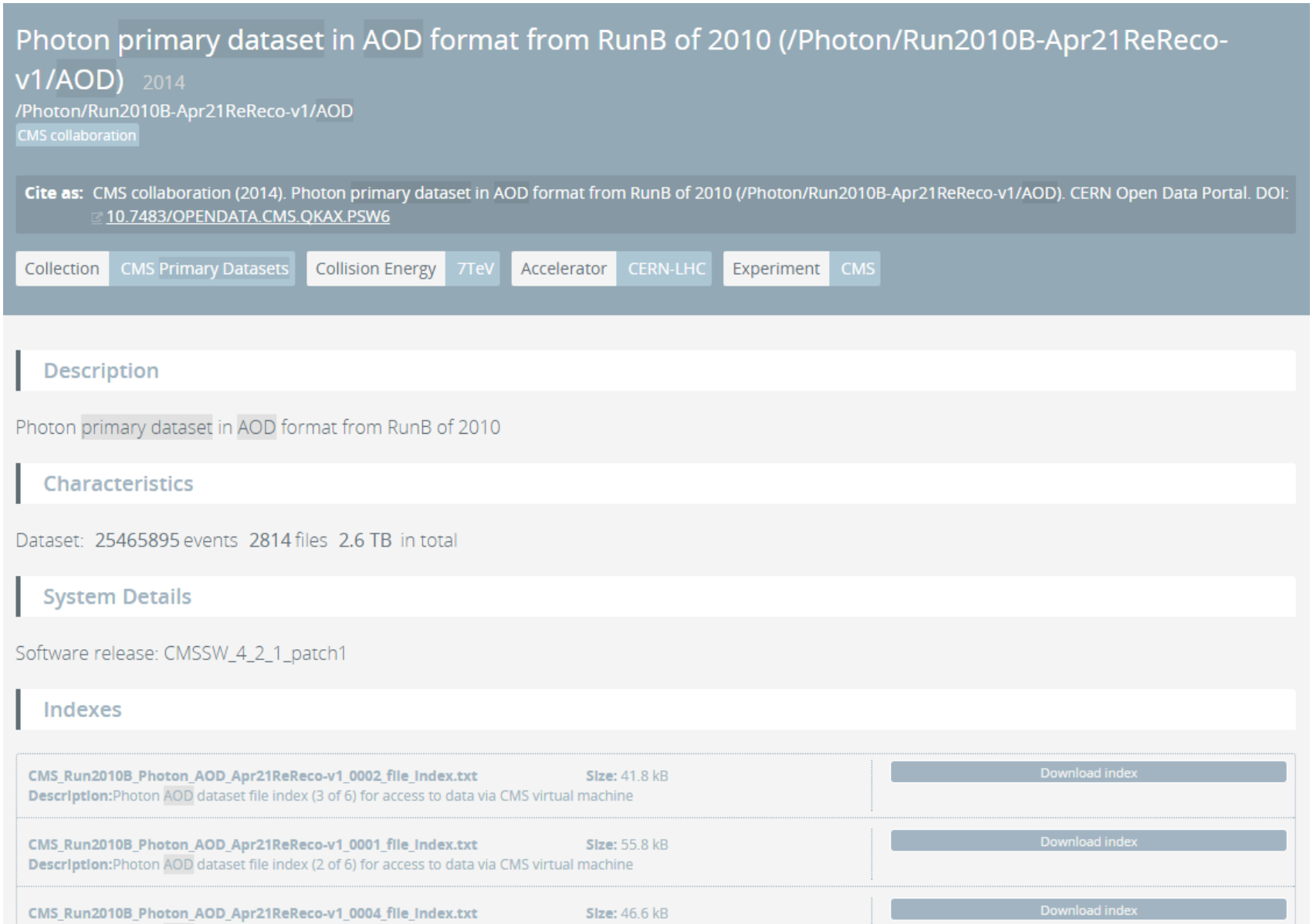
Visualise detector events



Basic histogramming



CMS primary datasets



Photon primary dataset in AOD format from RunB of 2010 (/Photon/Run2010B-Apr21ReReco-v1/AOD) 2014

/Photon/Run2010B-Apr21ReReco-v1/AOD

CMS collaboration

Cite as: CMS collaboration (2014). Photon primary dataset in AOD format from RunB of 2010 (/Photon/Run2010B-Apr21ReReco-v1/AOD). CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.QKAX.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

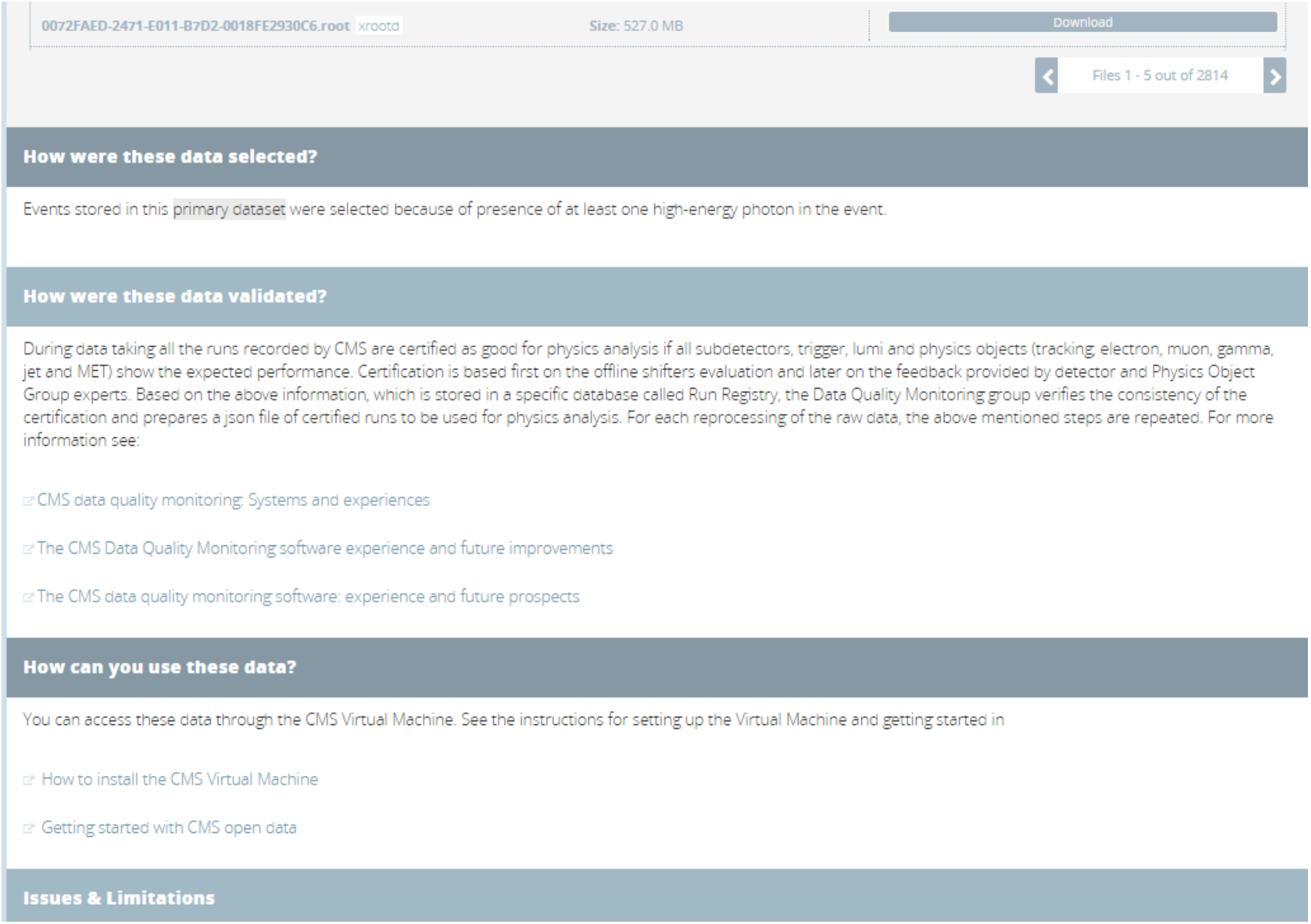
System Details

Software release: CMSSW_4_2_1_patch1

Indexes

CMS_Run2010B_Photon_AOD_Apr21ReReco-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		Download index
CMS_Run2010B_Photon_AOD_Apr21ReReco-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		Download index
CMS_Run2010B_Photon_AOD_Apr21ReReco-v1_0004_file_index.txt **Size:** 46.6 kB		Download index

CMS primary datasets



IPython notebooks

nbviewer FAQ IPython Jupyter

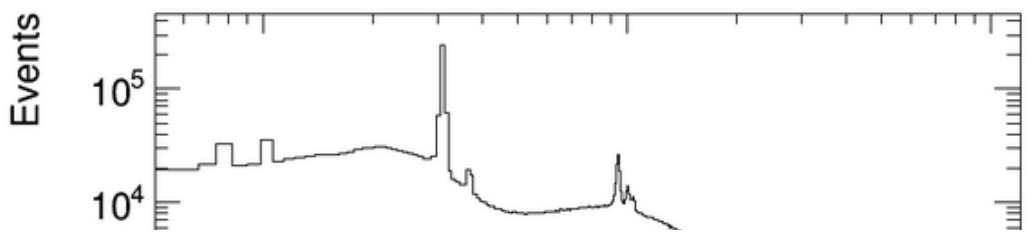
Write the analysis code

Loop on CMS dimuon events in a CSV file and calculate the dimuon invariant mass. These events were extracted from the CMS Mu Primary Dataset on the CERN Open Data Portal: http://opendata.cern.ch/record/14. Thanks to http://openstack.cern.ch for providing a CernVM running SL5 where I could set up CMSSW 4.2.8 and to https://github.com/tpmccauley/dimuon-filter for generating the CSV from CMS AODs.

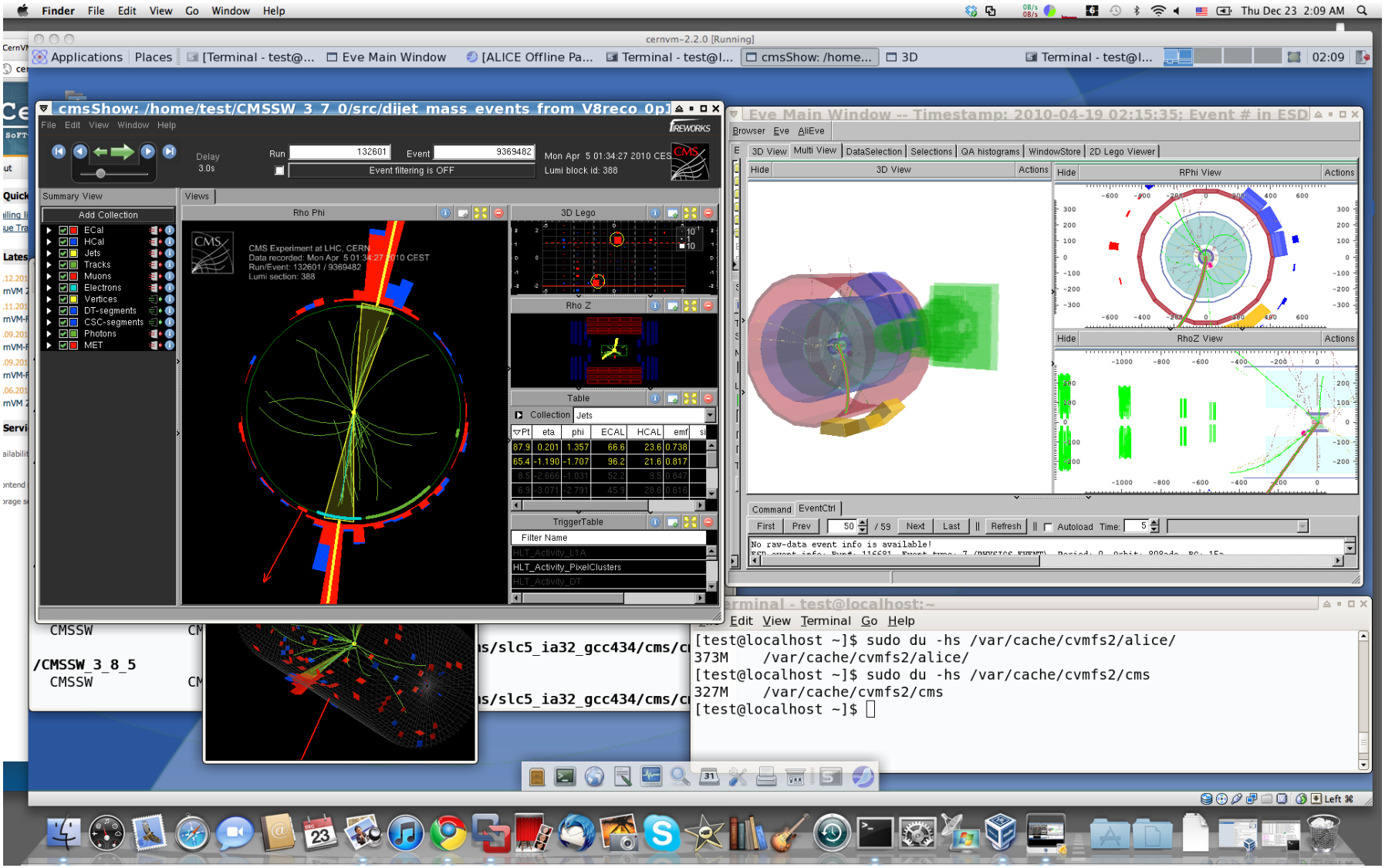
In [3]:
```
def analysis(filename):
    import csv
    from math import sqrt
    from rootpy.plotting import Hist
    from root_numpy import hist2array

    h = Hist(1500, 0.5, 120)

    with open(filename, 'rb') as csvfile:
        reader = csv.reader(csvfile, delimiter=',')
        header = reader.next()
        column = dict([(header[i], i) for i in range(len(header))])
```

In [8]:
```
plot(arrays, Hist(1500, 0.5, 120))
```

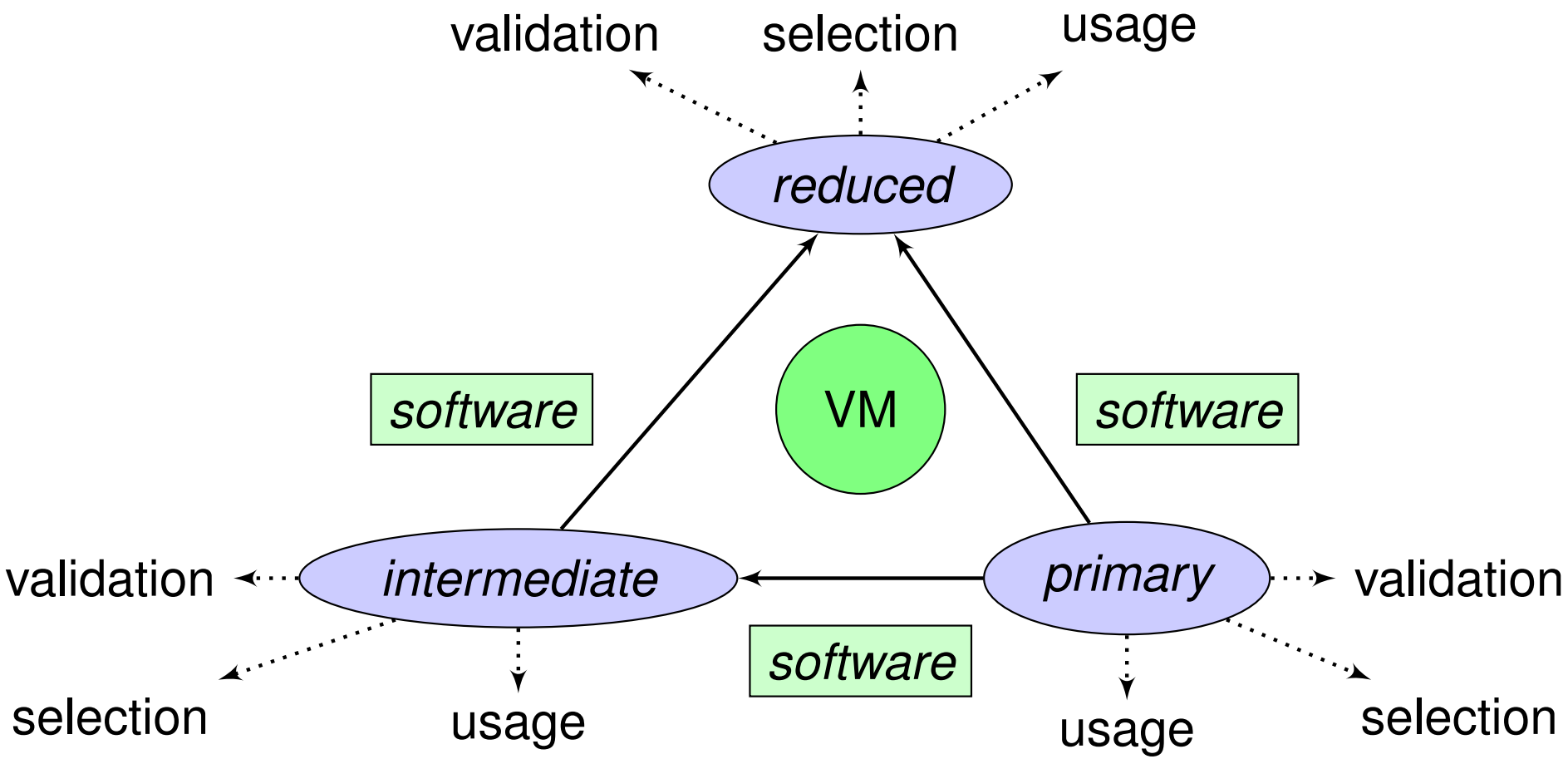
Out[8]:



CernVM virtual machines



OS ↔ Data ↔ Software

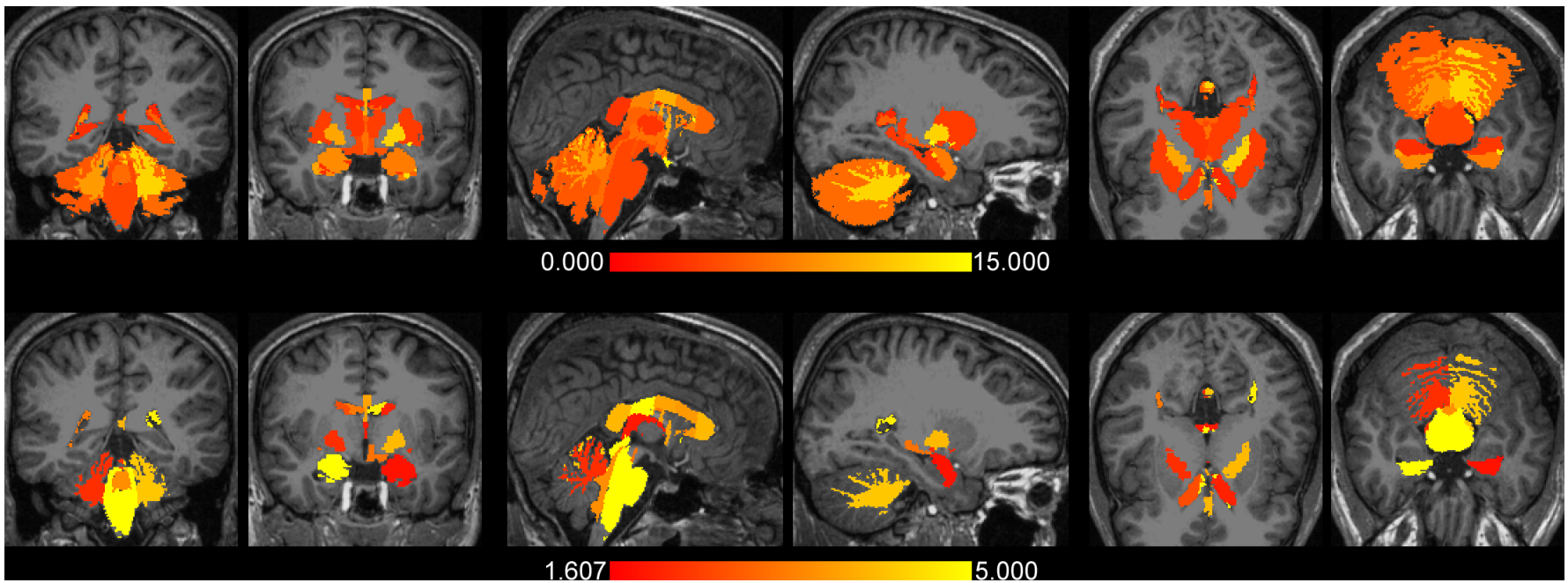


OS/Software influence in medecine

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

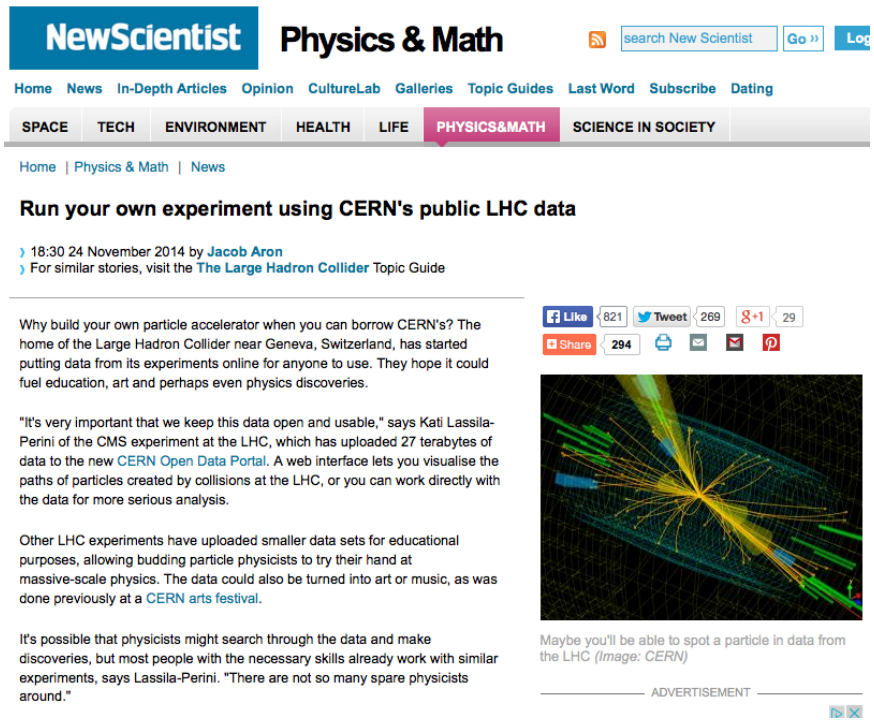
Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Published: June 1, 2012 • DOI: 10.1371/journal.pone.0038234



8.8±6.6% (volume) and 2.8±1.3% (cortical thickness)

Open data? Who cares?



NewScientist Physics & Math

search New Scientist Go »

Home News In-Depth Articles Opinion CultureLab Galleries Topic Guides Last Word Subscribe Dating

SPACE TECH ENVIRONMENT HEALTH LIFE PHYSICS&MATH SCIENCE IN SOCIETY

Home | Physics & Math | News

Run your own experiment using CERN's public LHC data

› 18:30 24 November 2014 by Jacob Aron
› For similar stories, visit the The Large Hadron Collider Topic Guide

Why build your own particle accelerator when you can borrow CERN's? The home of the Large Hadron Collider near Geneva, Switzerland, has started putting data from its experiments online for anyone to use. They hope it could fuel education, art and perhaps even physics discoveries.

"It's very important that we keep this data open and usable," says Kati Lassila-Perini of the CMS experiment at the LHC, which has uploaded 27 terabytes of data to the new CERN Open Data Portal. A web interface lets you visualise the paths of particles created by collisions at the LHC, or you can work directly with the data for more serious analysis.

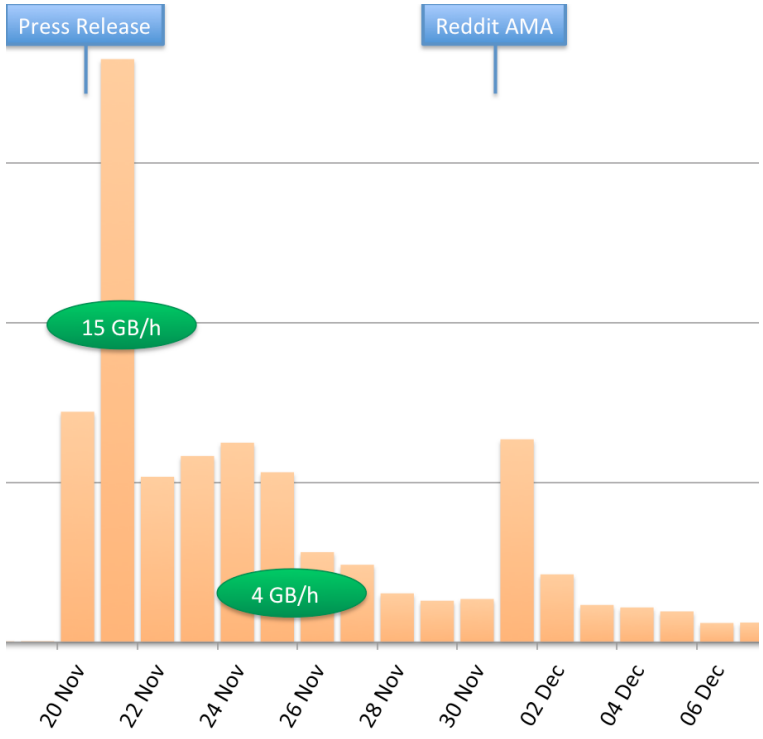
Other LHC experiments have uploaded smaller data sets for educational purposes, allowing budding particle physicists to try their hand at massive-scale physics. The data could also be turned into art or music, as was done previously at a CERN arts festival.

It's possible that physicists might search through the data and make discoveries, but most people with the necessary skills already work with similar experiments, says Lassila-Perini. "There are not so many spare physicists around."

Like 821 Tweet 269 +1 29 Share 294

Maybe you'll be able to spot a particle in data from the LHC *(Image: CERN)*

ADVERTISEMENT



82,000 distinct users visited the site

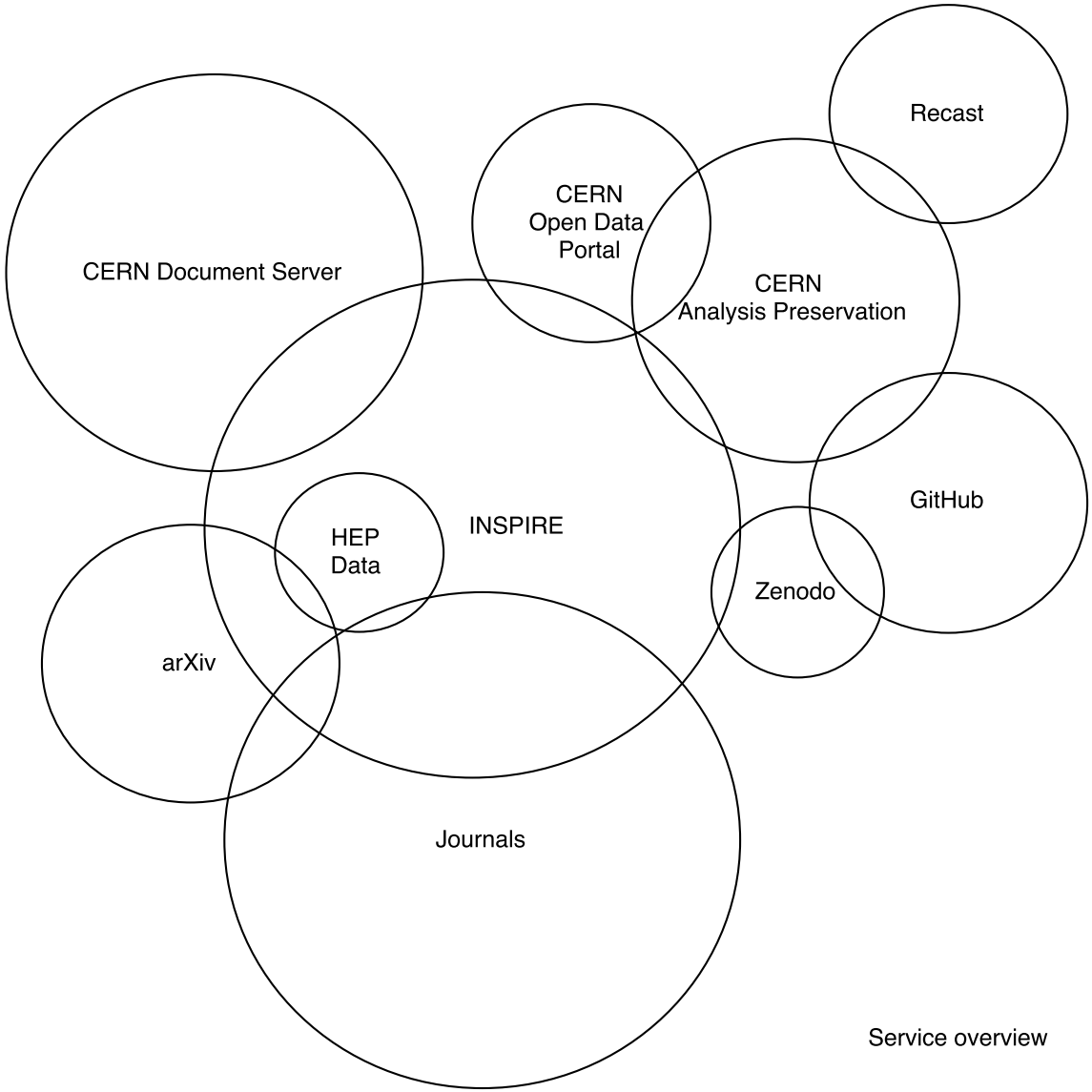
21,000 distinct users viewed data records

16,000 distinct users used event display

3,000 distinct users used histogramming

Conclusions

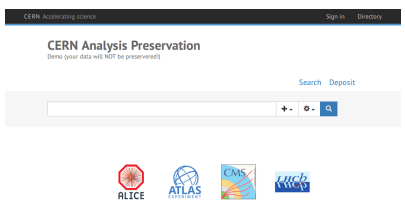
CERN (open) data services



CERN (open) data services

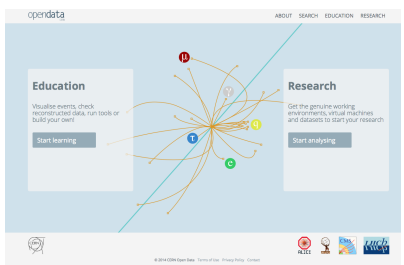
Invenio

http://inveniosoftware.org
http://github.com/inveniosoftware
@inveniosoftware
info@inveniosoftware.org



CERN Analysis Preservation

http://analysis-preservation.cern.ch
http://github.com/cernanalysispreservation
analysis-preservation-development@cern.ch



CERN Open Data

http://opendata.cern.ch
http://github.com/cernopendata
opendata-development@cern.ch

CERN IT J. Kunčar, D. Rodriguez, T. Šimko · **CERN Library** S. Dallmeier-Tiessen, R. Dasler, P. Fokianos, A. Lavasa, A. Mattmann, I. Tsanaktsidis, A. Trzcinska · **ALICE** M. Gheata, C. Grigoras · **ATLAS** K. Cranmer, L. Heinrich, A. Sanchez Pineda, D. Rousseau, F. Socher · **CMS** A. Calderon, A. Geiser, A. Huffman, K. Lassila-Perini, T. McCauley, A. Rao, A. Rodriguez Marrero · **LHCb** S. Amerio, B. Couturier, A. Trisovic · **CERN CernVM** J. Blomer · **CERN EOS** L. Mascetti · **DASPOS** M. Hildreth · **DPHEP** F. Berghaus, J. Shiers