

CERN Open Data and Data Analysis Knowledge Preservation

Tibor Šimko



Digital Library 2015 · 21–23 April 2015 · Jasná, Slovakia

1

Invenio

What is Invenio?

- digital library and document repository software
 - mature platform: first public release in 2002
 - rich data: articles, books, notes, photos, videos, *software*, *data*
- some Invenio-based services at CERN:

CERN Document Server

Access articles, reports and multimedia content in HEP



- co-developed by an international collaboration



- participating in EU FP7/H2020 projects



“Data behind plots”



Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNames :: Institutions :: Conferences :: Jobs :: Experiments :: Journals :: Help

Information References Citations **Plots** HepData

Search for new phenomena in final states with large jet multiplicities and missing transverse momentum at $\sqrt{s} = 8$ TeV proton-proton collisions using the ATLAS experiment - ATLAS Collaboration (Aad, Georges et al.) JHEP 1310 (2013) 130 arXiv:1308.1841 [hep-ex] CERN-PH-EP-2013-110

THIS DATA COMES FROM [DURHAM HEPDATA PROJECT](#)

DATASETS:

Description: MET/sqrt(HT) distributions for the multi-jet + flavour stream with PTmin=50 GeV, and exactly eight jets, with the signal region selection, other than that on MET/sqrt(HT) itself. A selection of zero b-jets is applied.

[Go to the record](#)

Plain

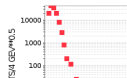
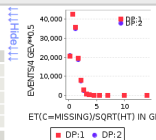
$|ETARAP(B^{-}jet)| < 2.5$
 $N - BJETS(p_T > 40 \text{ GeV}) = 0.0$
 $N - JETS(p_T > 50 \text{ GeV}) = 8.0$
 $pp \rightarrow JETS MM$
 $= dATA = MC$
 $|ETARAP(jet)| < 2.0$

$ET_{MISSING}/\sqrt{HT} \text{ (GeV}^{0.5}\text{)}$

$EVENTS/4 \text{ GeV}^{0.5}$

+++Collapse+++

0.25	20504.0	21017.74 ± 0.25
0.75	42632.0	42806.83 ± 0.00
1.25	35848.0	35159.43 ± 0.00
1.75	19376.0	18926.27 ± 0.00
2.25	7872.0	7742.63 ± 0.00
2.75	2720.0	2686.48 ± 1.43
3.25	792.0	880.64 ± 4.83
3.75	100.0	200.77 ± 0.73



“Code you can cite”

- automated GitHub ↔ Zenodo bridge
- push new release to GitHub → automatic archival on Zenodo
- software preserved, minted with a DOI, made citable

The image is a collage of four screenshots illustrating the Zenodo-GitHub integration process:

- Top Left:** The Zenodo website's 'Get started' guide for GitHub repositories. It shows steps: 1. Flip the switch, 2. Create a release, 3. Set the badge. A blue arrow points from the 'Flip the switch' step to the 'ON' toggle in the bottom right.
- Top Right:** A GitHub repository page for 'invenio-coudecouple'. It shows the 'Releases' tab with a list of releases. A blue arrow points from the 'v1.1.3' release to the 'Releases' section in the bottom right.
- Bottom Left:** A screenshot of the GitHub Releases page for 'invenio-coudecouple' v1.1.3. It shows the release details, including the source code and a download link. A blue arrow points from the 'v1.1.3' release to the 'Releases' section in the bottom right.
- Bottom Right:** A screenshot of the '.zenodo.json' file. It shows the JSON structure for the release, including the name, description, and related identifiers. A blue arrow points from the 'v1.1.3' release to the 'Releases' section in the bottom right.

Arrows indicate the flow of information and actions between these components, showing how a release on GitHub is automatically archived on Zenodo and minted with a DOI.

<https://guides.github.com/activities/citable-code>

- link data (DATAVERSE) to code (ZENODO) to papers (INSPIRE)
- example: hep-ex/0011057, arXiv:1401.0080

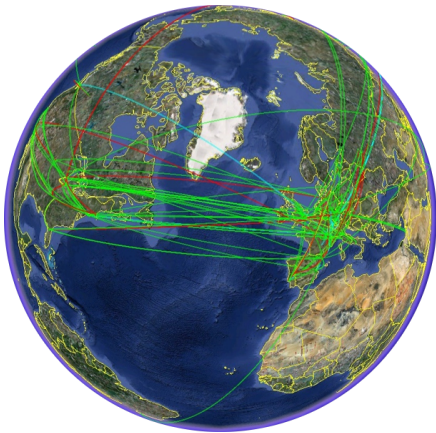
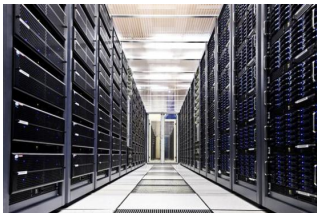
The figure displays three overlapping screenshots from the INSPIRE database, illustrating the integration of software and publications. The leftmost screenshot shows a search result for 'decouple software associated to arXiv:1401.0080'. The middle screenshot provides a detailed view of this software entry, including its description and a link to the associated publication. The rightmost screenshot shows the publication entry for 'A Novel Approach to Higgs Coupling Measurements', highlighting the software's role in the research.

2

Data Analysis

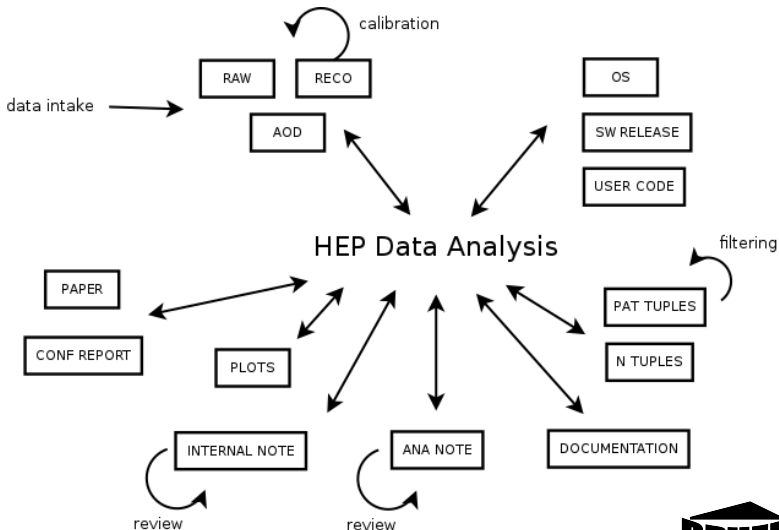


Large Scale Solutions



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

Preserve an Analysis?



Big Data?

data

raw

reconstructed

reduced

publication

scale

~GB / sec

~PB / year

~TB / analysis

~GB / analysis

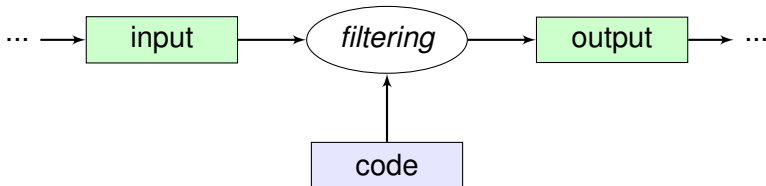
knowledge

calibration, conditioning

filtering, selection

user code, physics objects

correlation, data behind plots



Analysis Train

Final selection step



OS

Analysis software

User code

Example of supported repositories:

git@github.com:johndoe/myrepo.git

svn@svnweb.cern.ch/cern/wsvn/myrepo

Harvest? ☒ yes, harvest user code

☐ no, create link only

Input Data Files ☒ taken from output of pre-selection step

☐ taken from output of custom analysis step

Output Data Files [+ Add another file](#)

Example of supported protocols:

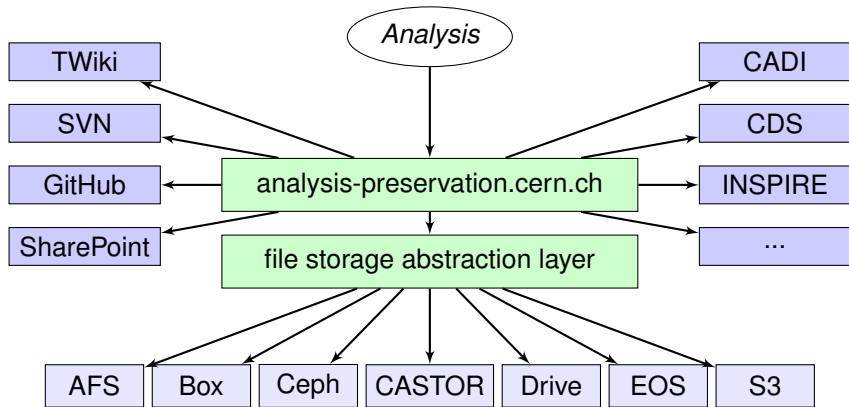
xroot://castorpublic.cern.ch/castor/cern.ch/user/j/johndoe/mydir/myfile.root

root://eospublic.cern.ch/eos/lhcb/.../myfile.root

file:///tmp/myfile.root

http://john.doe.example.org/myfile.root

Harvest? ☒ yes, harvest files ☐ no, create link only



■ record format: extended MARC21

- “technical” metadata: beyond bytes

e.g. 256 “computer file characteristics”

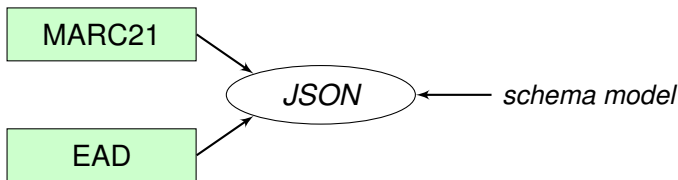
\$a characteristics	\$e events	\$t text
\$b bytes	\$f files	...

- “knowledge” metadata: semantics

e.g. 505 “formatted contents note” CSV column information

\$t title	\$g miscellaneous
-----------	-------------------

■ internal format: JSON



3

Open Data

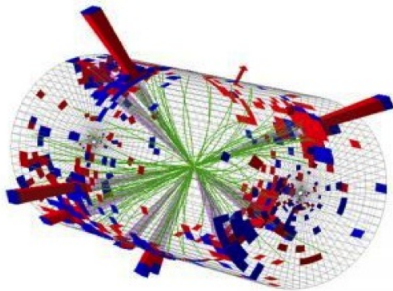
Data policies:

- restricted → embargo period → 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:
 - data miners
 - citizen scientists
 - high-school students
 - general public
- computing:
 - exploring in the browser
 - specialised VMs



opendata

ABOUT SEARCH EDUCATION RESEARCH

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

© 2014 CERN Open Data Terms of Use Privacy Policy Contact

ALICE ATLAS CMS LHCb

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 >](#)

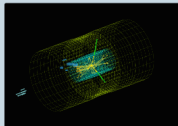
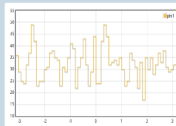
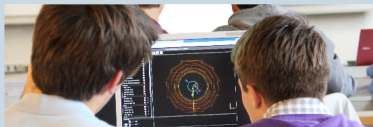
The ATLAS (A Toroidal LHC ApparatuS) experiment is a general purpose detector exploring topics like the properties of the Higgs-like particle, extra dimensions of space, unification of fundamental forces, and evidence for dark matter candidates in the Universe.

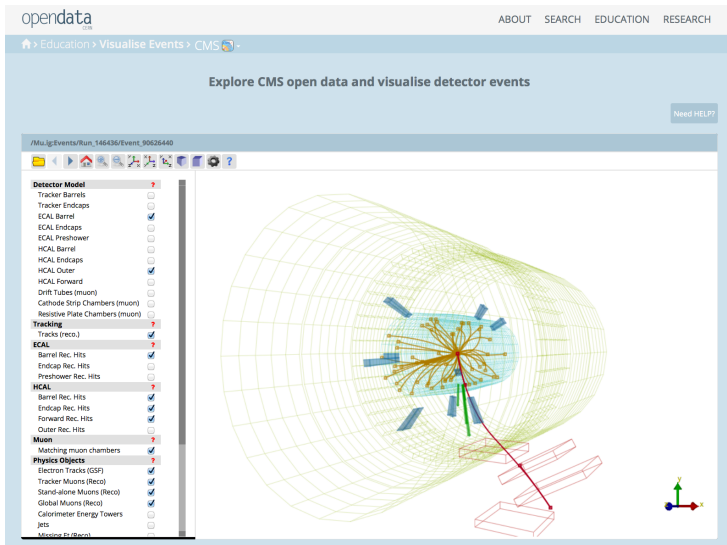
[Explore ATLAS >](#)

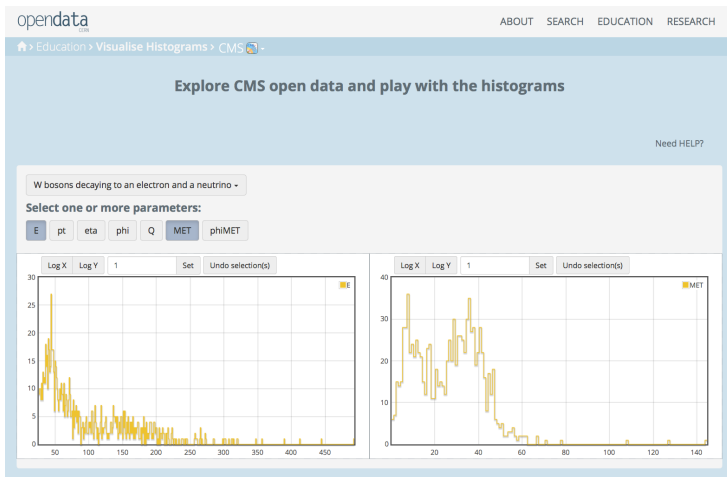
The LHCb (Large Hadron Collider beauty) experiment aims to record the decay of particles containing b and anti-b quarks, known as B mesons. The detector is designed to gather information about the identity, trajectory, momentum and energy of each particle.

[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 >](#)

opendata

ABOUT SEARCH EDUCATION RESEARCH

Q Search

Home > CMS > CMS Primary Datasets

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

/Mu/Run2010B-Apr21ReReco-v1/AOD

CMS collaboration

Cite as: CMS collaboration (2014). Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD). CERN Open Data Portal. DOI: [10.7483/OPENDATA.CMS.B8MR.C4A2](https://doi.org/10.7483/OPENDATA.CMS.B8MR.C4A2)

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

Description

Mu primary dataset in AOD format from RunB of 2010

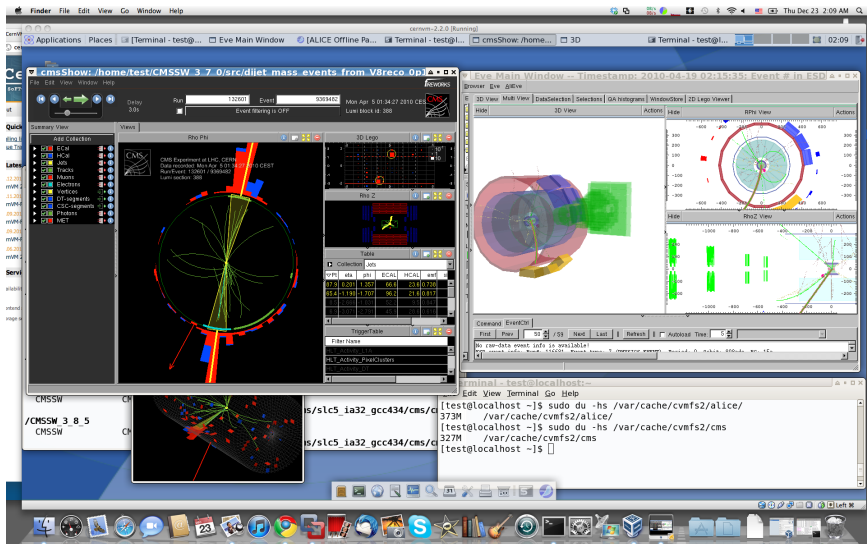
Characteristics

Dataset: 32376291 events 2979 files 3.2 TB in total

System Details

Software release: CMSSW_4_2_1_patch1

Indexes



Open Data? Who cares?



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

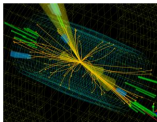
Why build your own particle accelerator when you can borrow CERN's? The home of the Large Hadron Collider near Geneva, Switzerland, has started putting data from its experiments online for anyone to use. They hope it could fuel education, art and perhaps even physics discoveries.

"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](#).

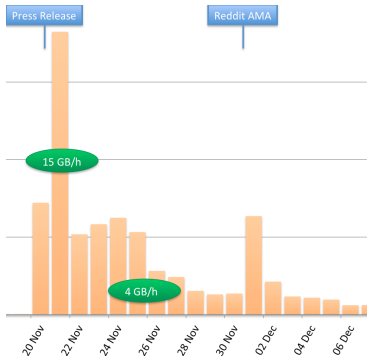
It's possible that physicists might search through the data and make discoveries, but most people with the necessary skills already work with similar experiments, says Lassila-Perini. "There are not so many spare physicists around."

Facebook 621, Twitter 269, Google+ 29, Dribbble 294, LinkedIn 294, YouTube 294, SoundCloud 294, DeviantArt 294



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

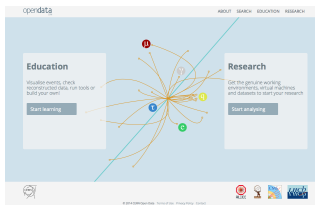
ADVERTISEMENT



- 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

7

Conclusions



*Capturing and disseminating knowledge
of data, code, platform, processes
to enable future data reuse*

(Open) Data Analysis Preservation Framework

<http://opendata.cern.ch/>

CERN IT J. Cowton, P. Fokianos, J. Kunčar, T. Smith, T. Šimko

CERN Library S. Dallmeier-Tiessen, P. Herterich, 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, B. Couturier, A. Trisovic

CERN CernVM J. Blomer

CERN EOS L. Mascetti

DASPOS M. Hildreth

DPHEP F. Berghaus