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

INSPIRE HEP

zenodo

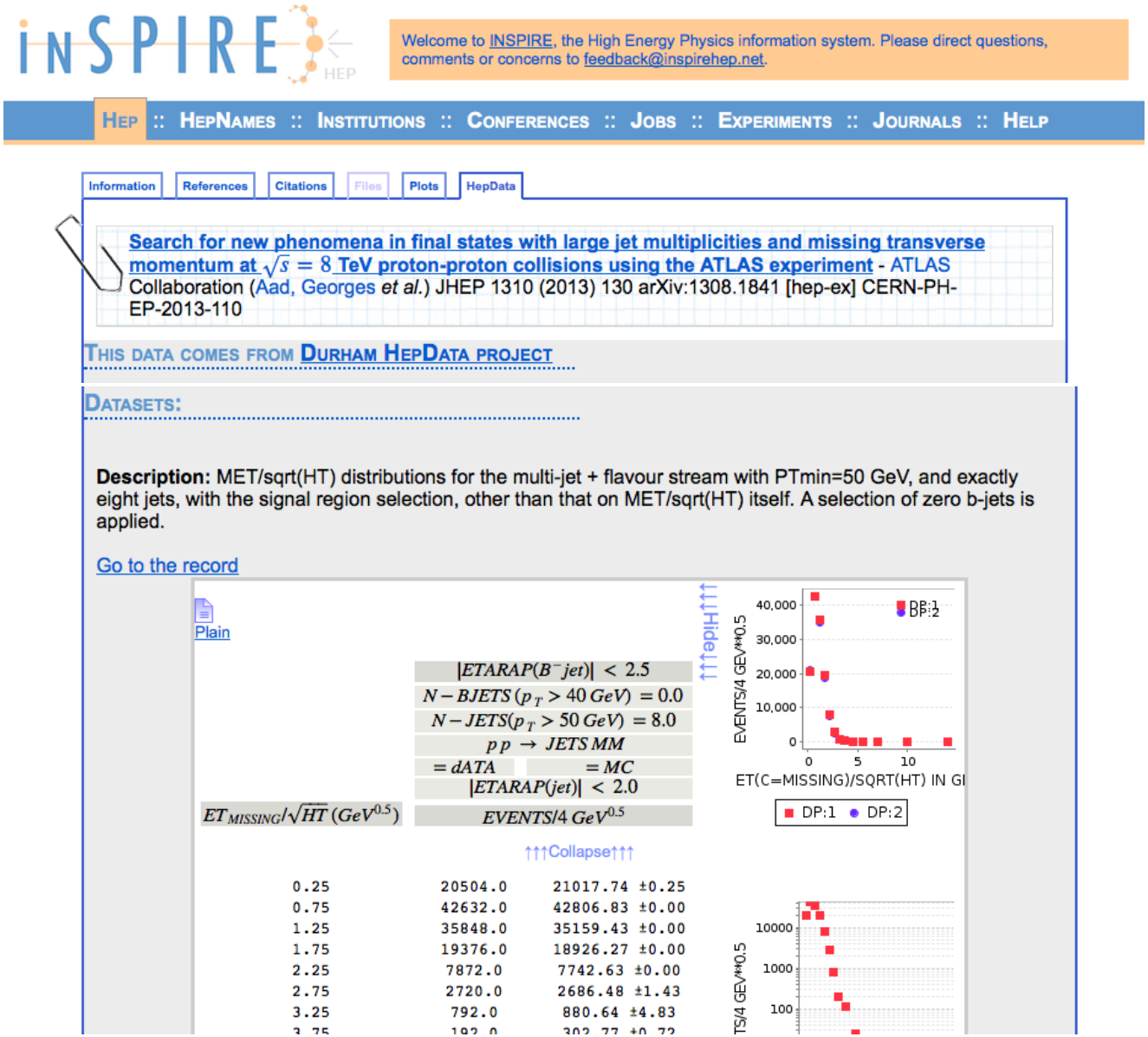
- co-developed by an international collaboration

CERN · DESY · EPFL École Polytechnique Fédérale de Lausanne · Fermilab · SLAC National Accelerator Laboratory · Harvard · International Labour Organization · JÜLICH Forschungszentrum · GSI · UAB Universitat Autònoma de Barcelona · Universidad Zaragoza · réro

- participating in EU FP7/H2020 projects

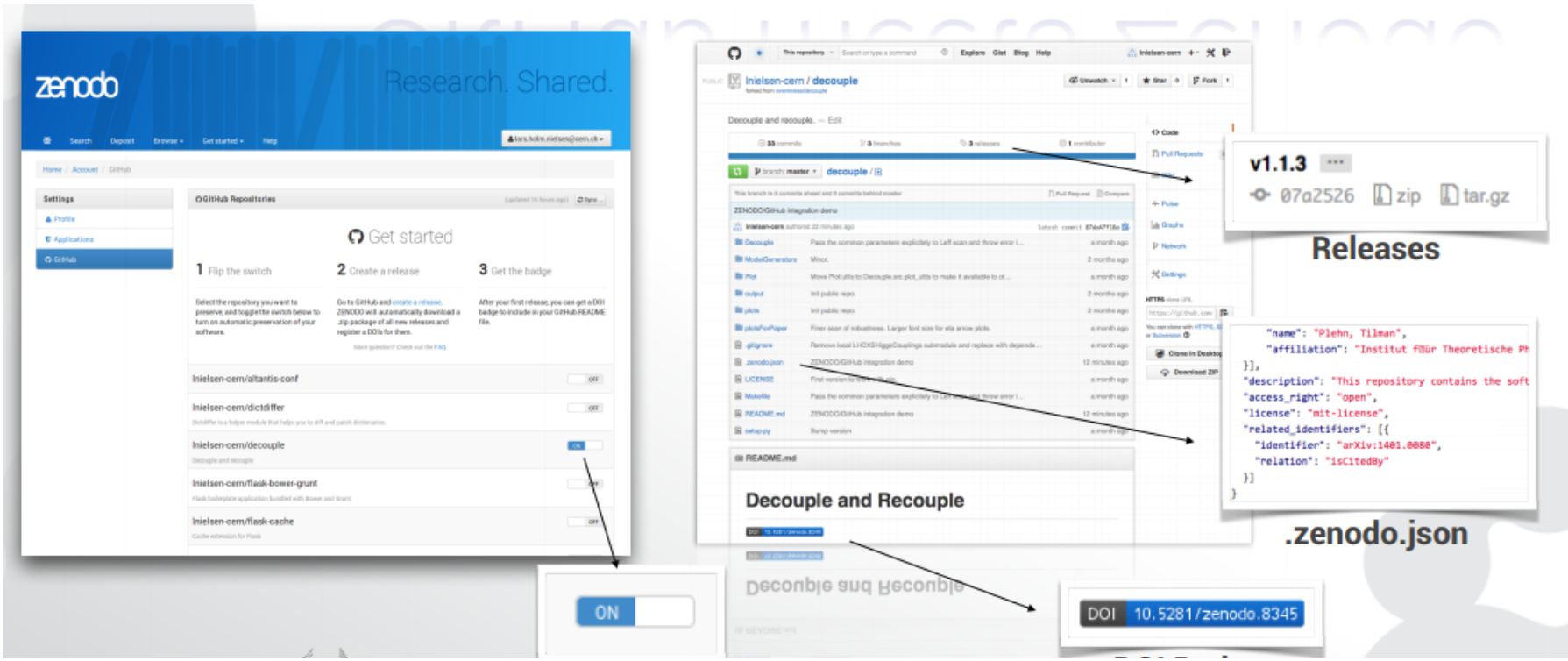
blog forever · CRISP The Cluster of Research Infrastructures for Synergies in Physics · EUDAT · ODIN ORCID and DataCite Interoperability Network · OpenAIRE Open Access Infrastructure for Research in Europe

“Data behind plots”



“Code you can cite”

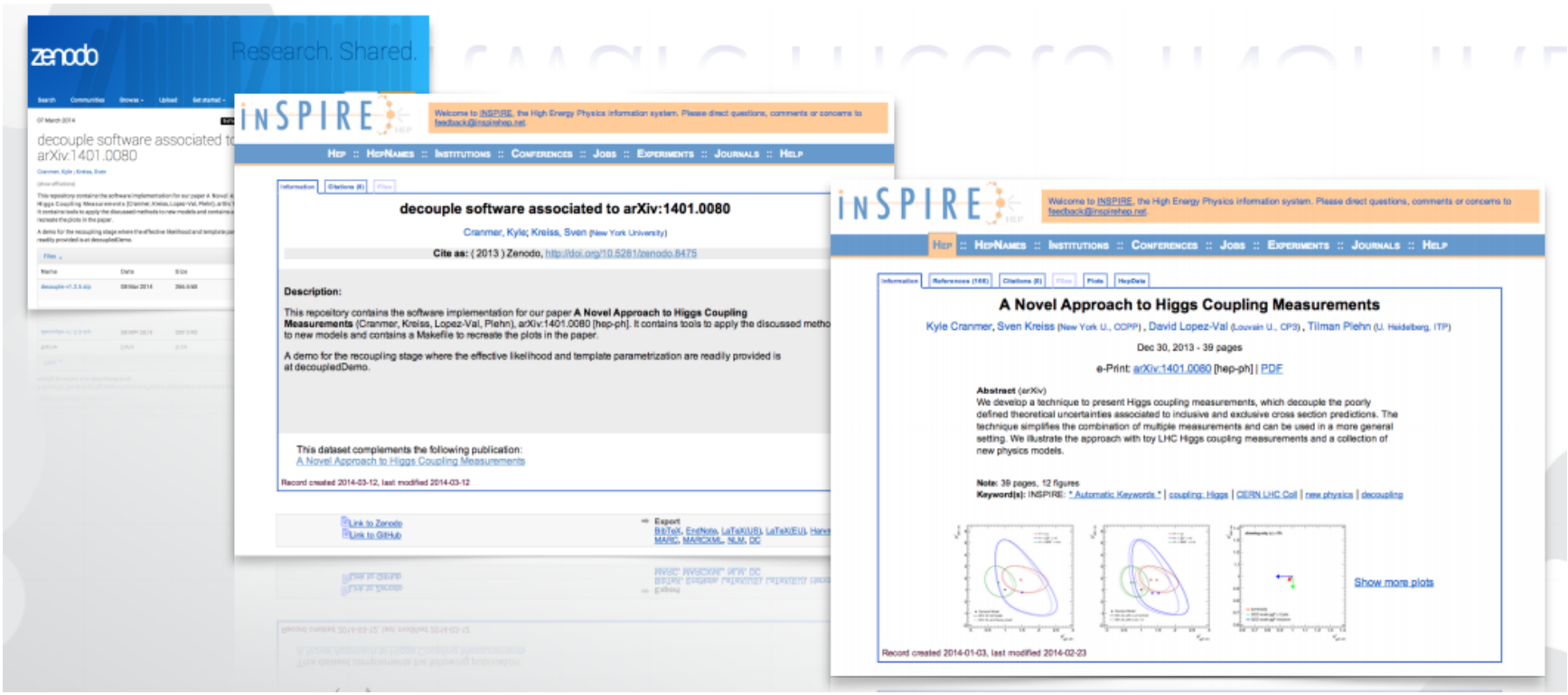
- automated GitHub ↔ Zenodo bridge
- push new release to GitHub → automatic archival on Zenodo
- software preserved, minted with a DOI, made citable



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

Code ↔ Data ↔ Paper

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

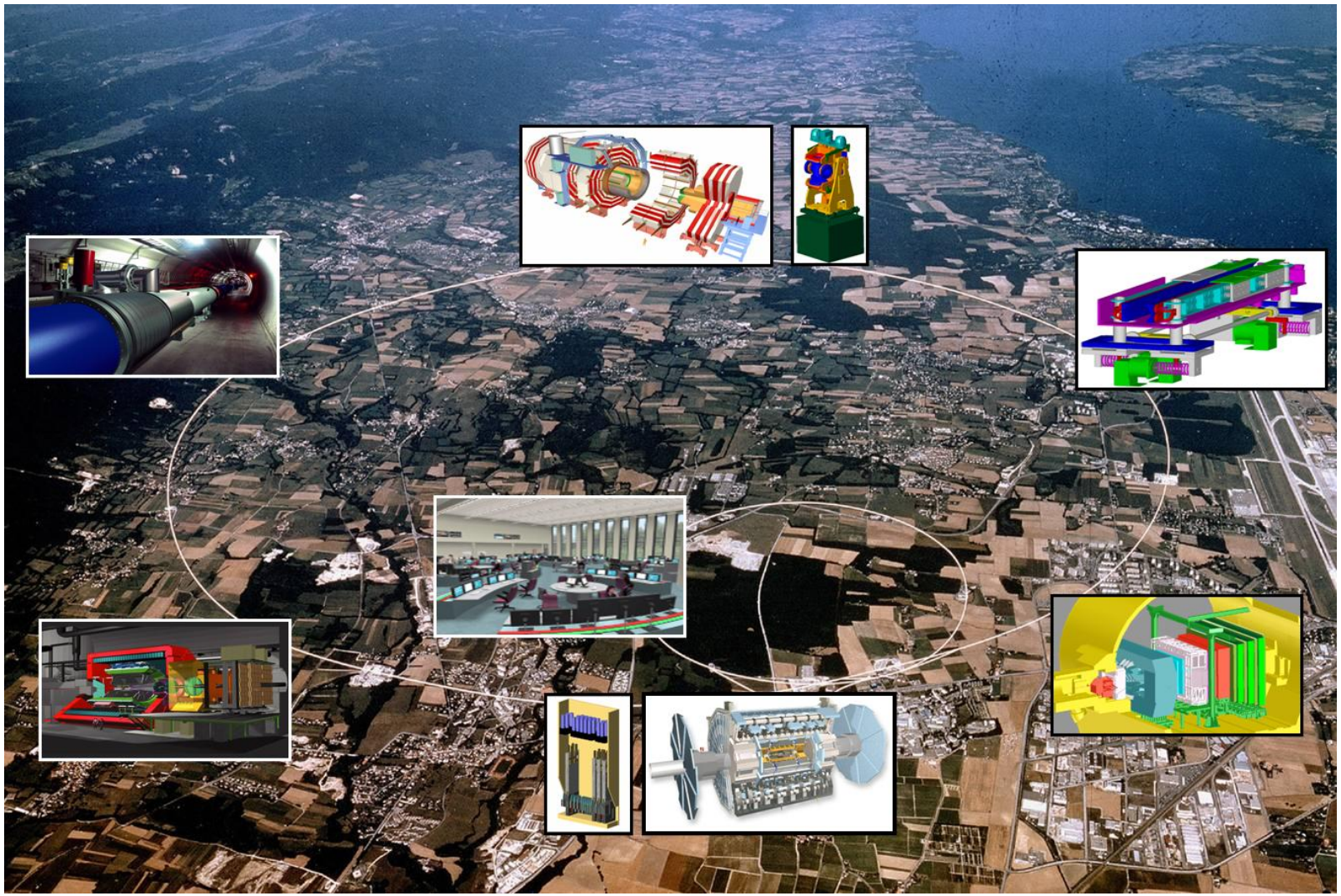


INVENIO

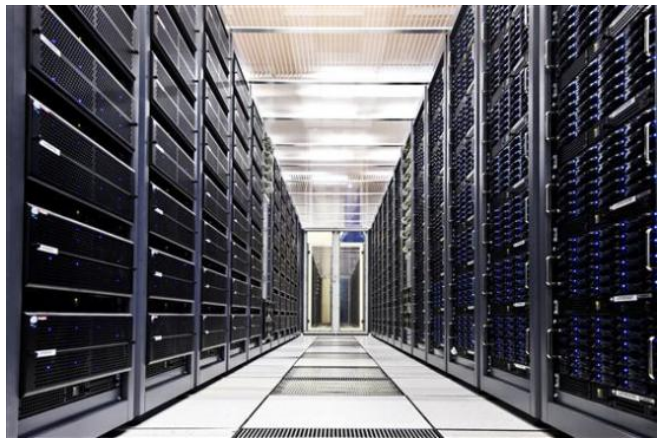
2

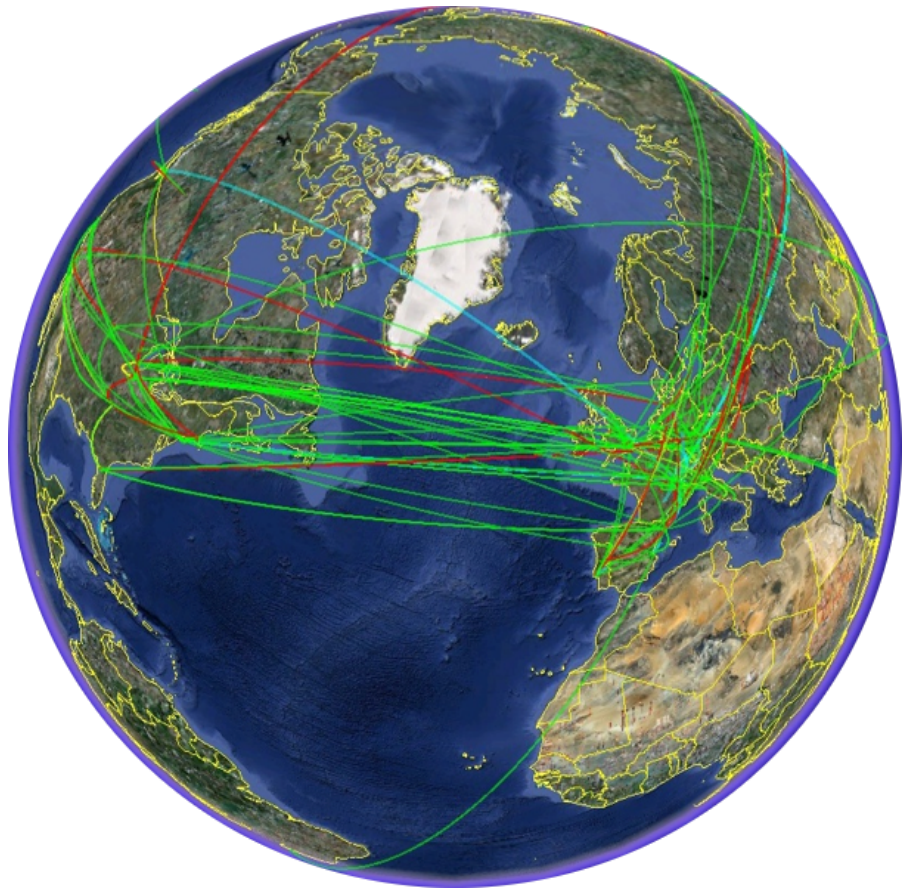
Data Analysis

CERN LHC Experiments



Large Scale Solutions

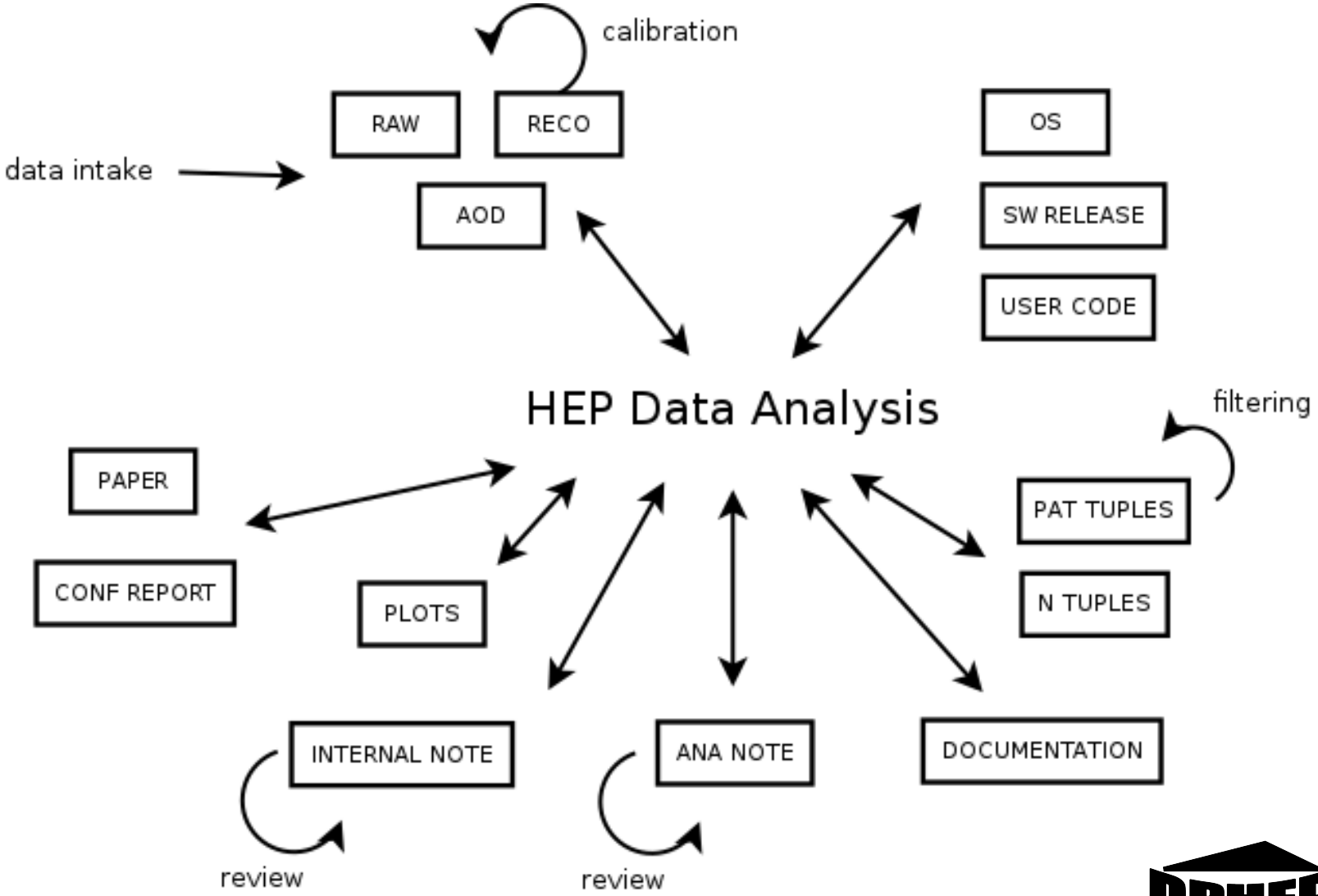




Primary site: 100k cores (10k nodes), 100k disks (50 PB), 21k NIC

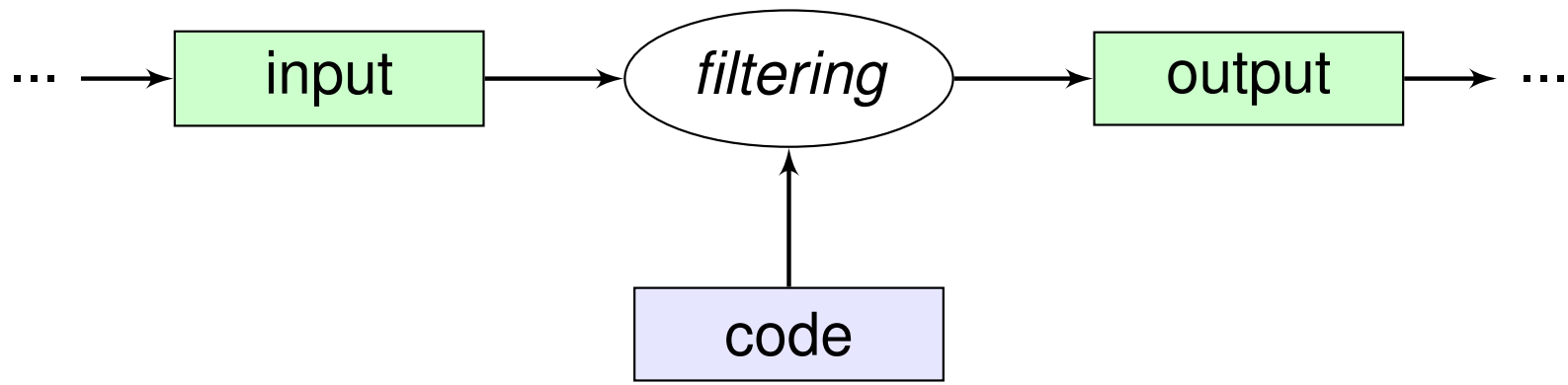
Grid: 13 Tier-1 sites, 155 Tier-2 sites, 10 Gbps optical fibre

Preserve an Analysis?



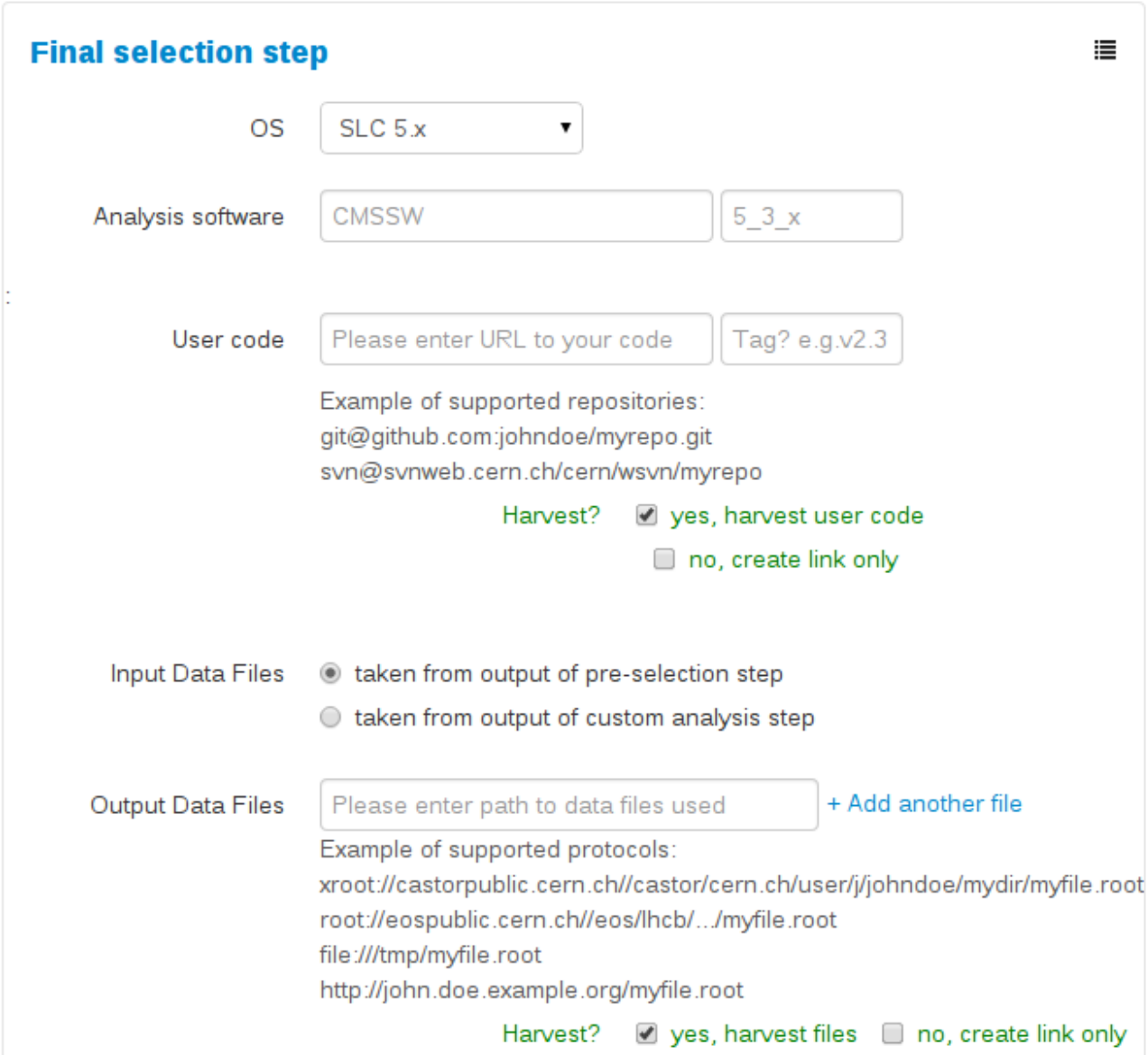
Big Data?

data	**scale**	**knowledge**
raw	∼GB / sec	calibration, conditioning
reconstructed	∼PB / year	filtering, selection
reduced	∼TB / analysis	user code, physics objects
publication	∼GB / analysis	correlation, data behind plots

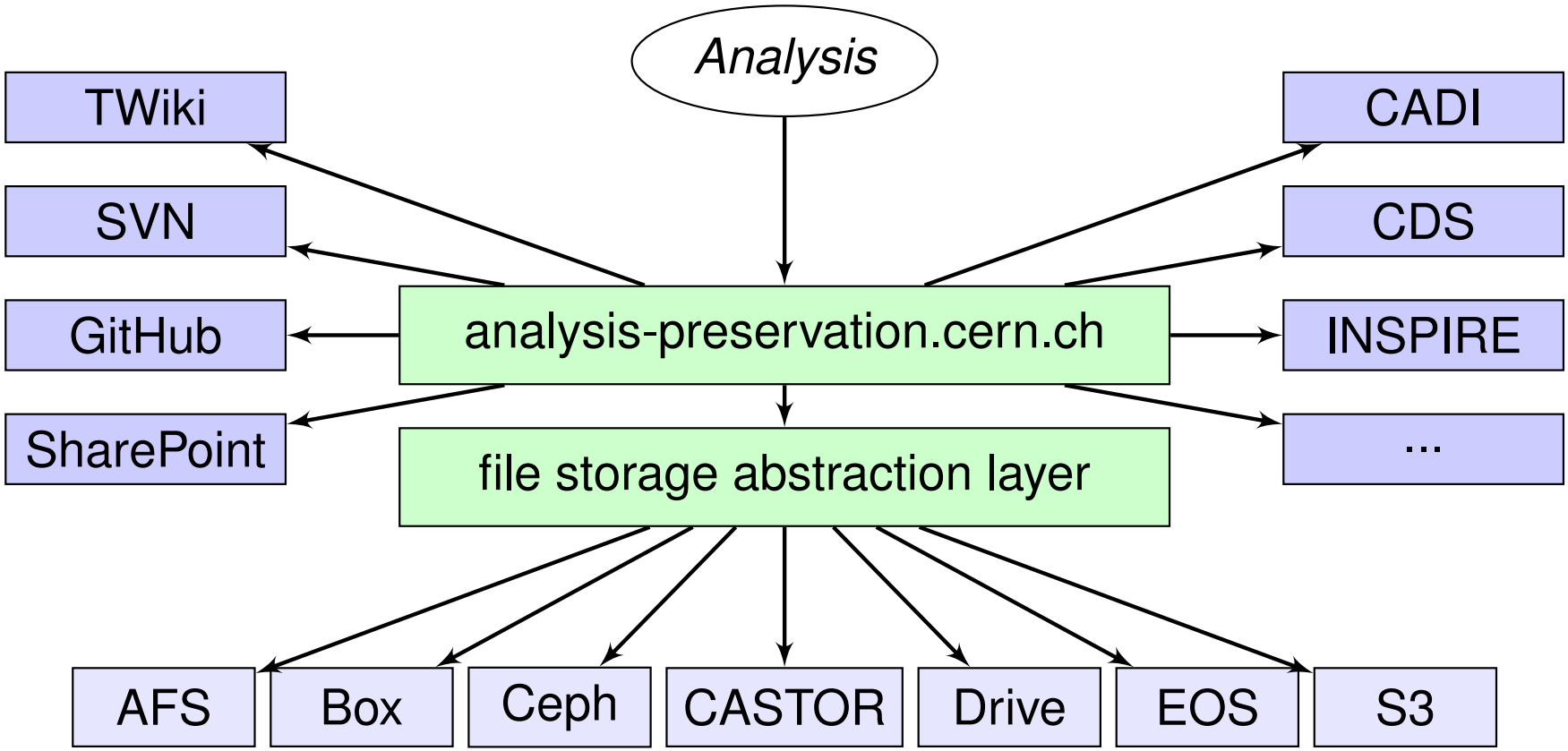


Analysis Train

Knowledge Capture



System Architecture

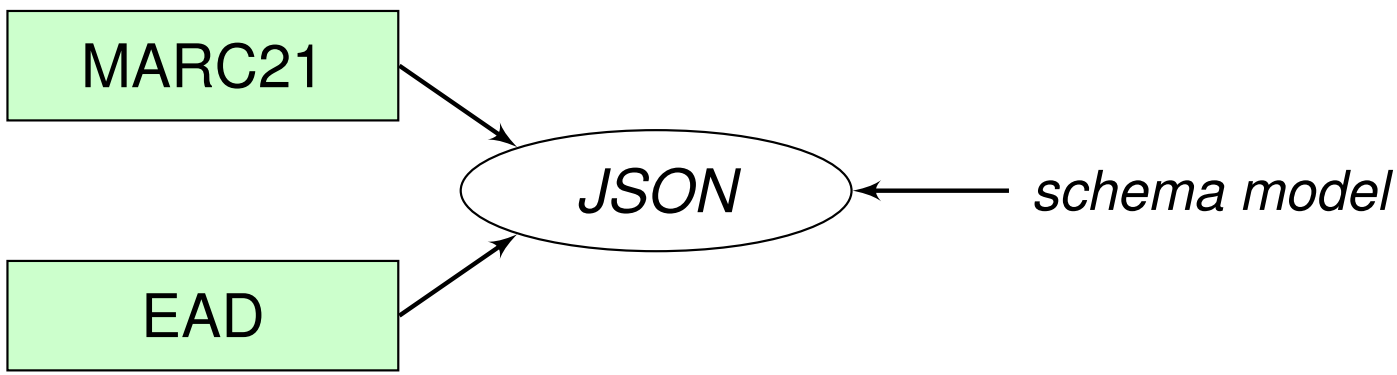


Knowledge Representation

- 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



INVENIO

3

Open Data

Opening Up

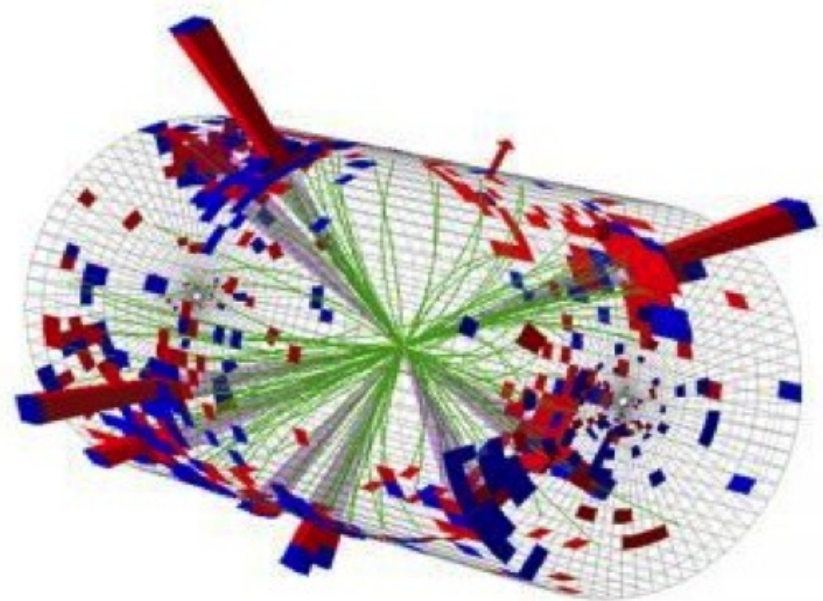
Data policies:

- restricted → embargo period → open

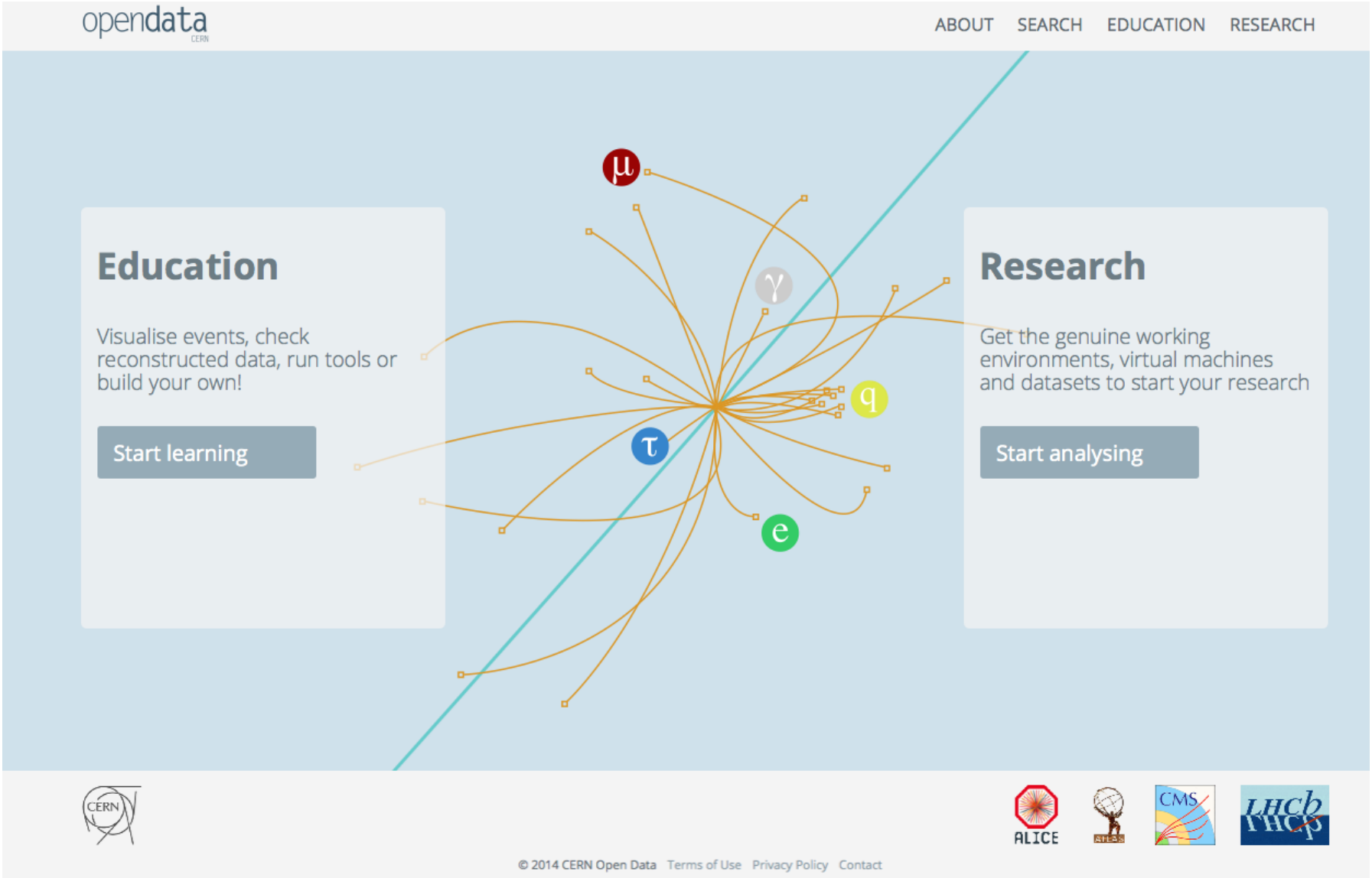
  "[...] Data with high abstraction, such as AOD, will be conditionally made publicly available after an embargo period of 5 years after publication for 10% of the data and 10 years for 100% of the data [...]" —ALICE Data Policy

Challenges:

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



CERN Open Data Portal



Education

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 >

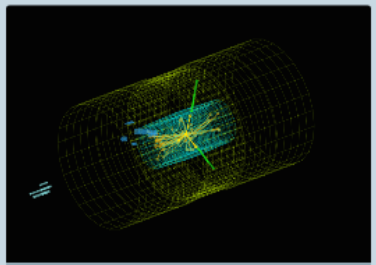
The ATLAS (A Toroidal LHC ApparatuS) experiment is a general purpose detector exploring topics like the properties of the Higgs-like particle, extra dimensions of space, unification of fundamental forces, and evidence for dark matter candidates in the Universe.

Explore ATLAS >

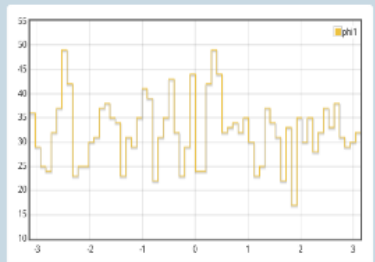
The LHCb (Large Hadron Collider beauty) experiment aims to record the decay of particles containing b and anti-b quarks, known as B mesons. The detector is designed to gather information about the identity, trajectory, momentum and energy of each particle.

Explore LHCb >

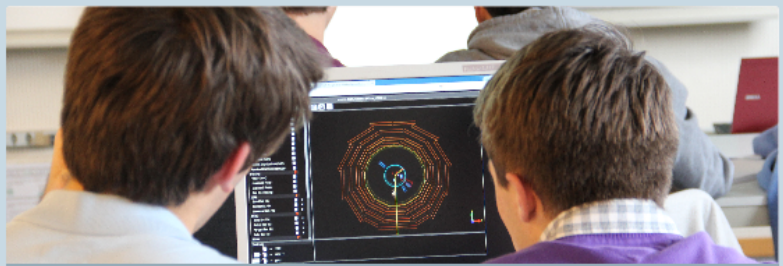
For education purposes, the complex primary data need to be processed into a format (examples below) that is good for simple applications. Get in touch if you wish to build your own applications similar to those shown here



Visualise events >

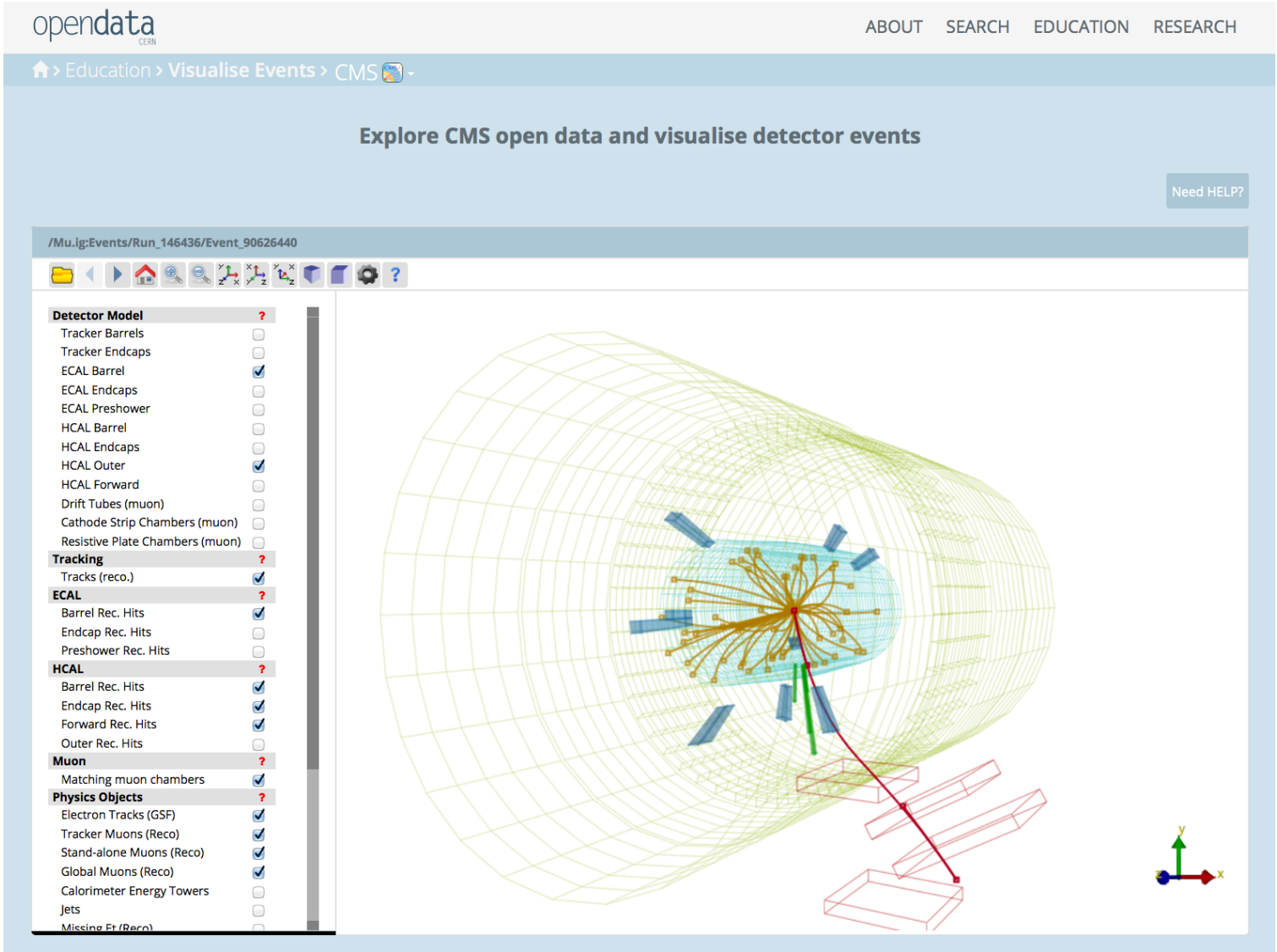


Visualise histograms >

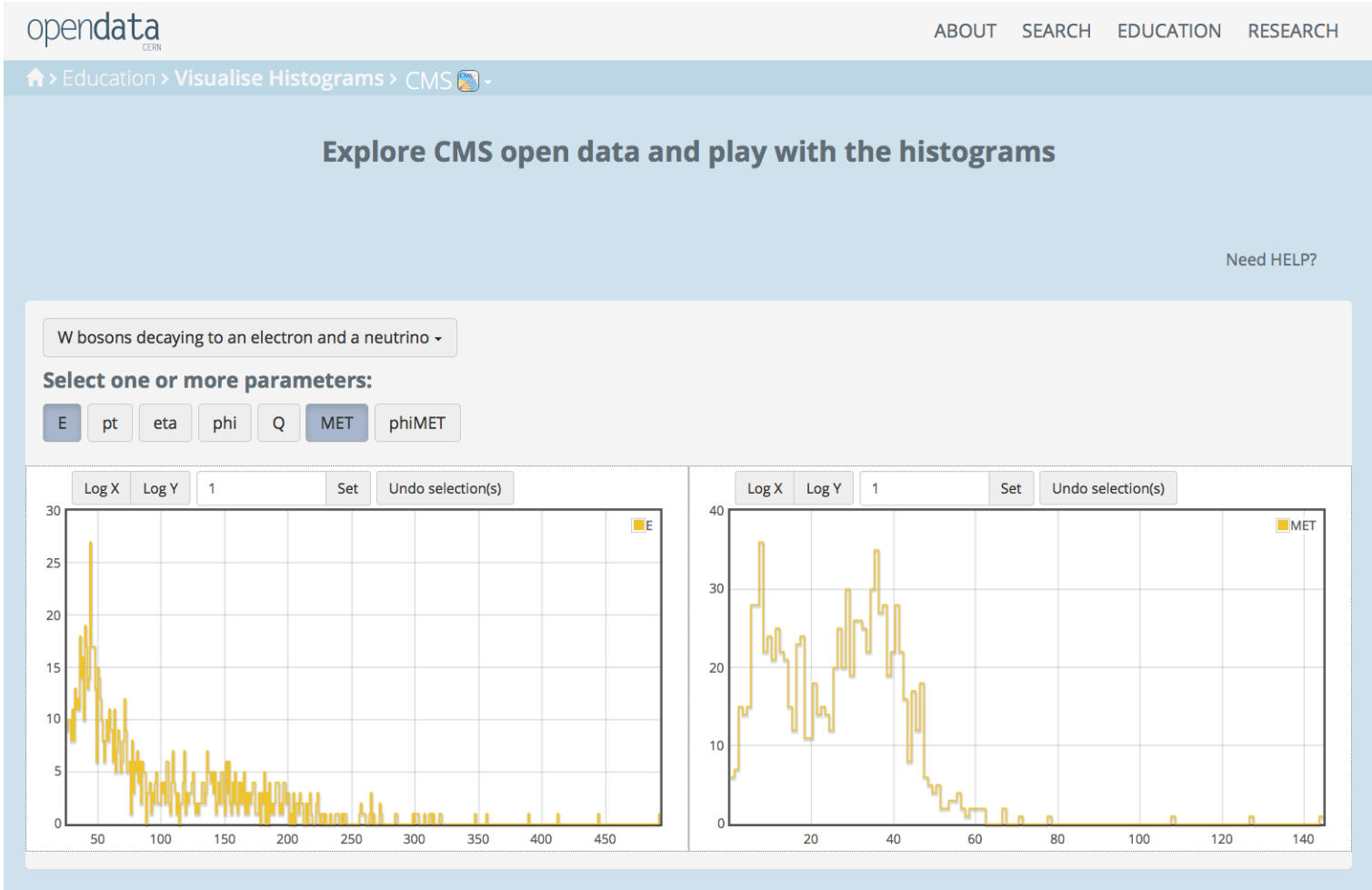


Learning Resources >

Visualise detector events



Basic histogramming



Research

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 >

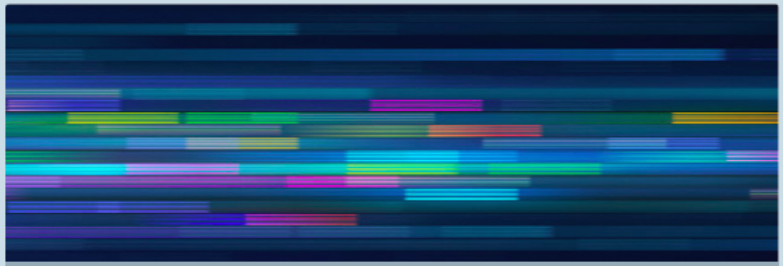
According to the ALICE data preservation strategy, reconstructed data and Monte Carlo data as well as the analysis software and documentation needed to process them will be made available on a time scale of 5 years (for 10% of the data). Thus, the first release of ALICE research data will happen in 2018.

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.

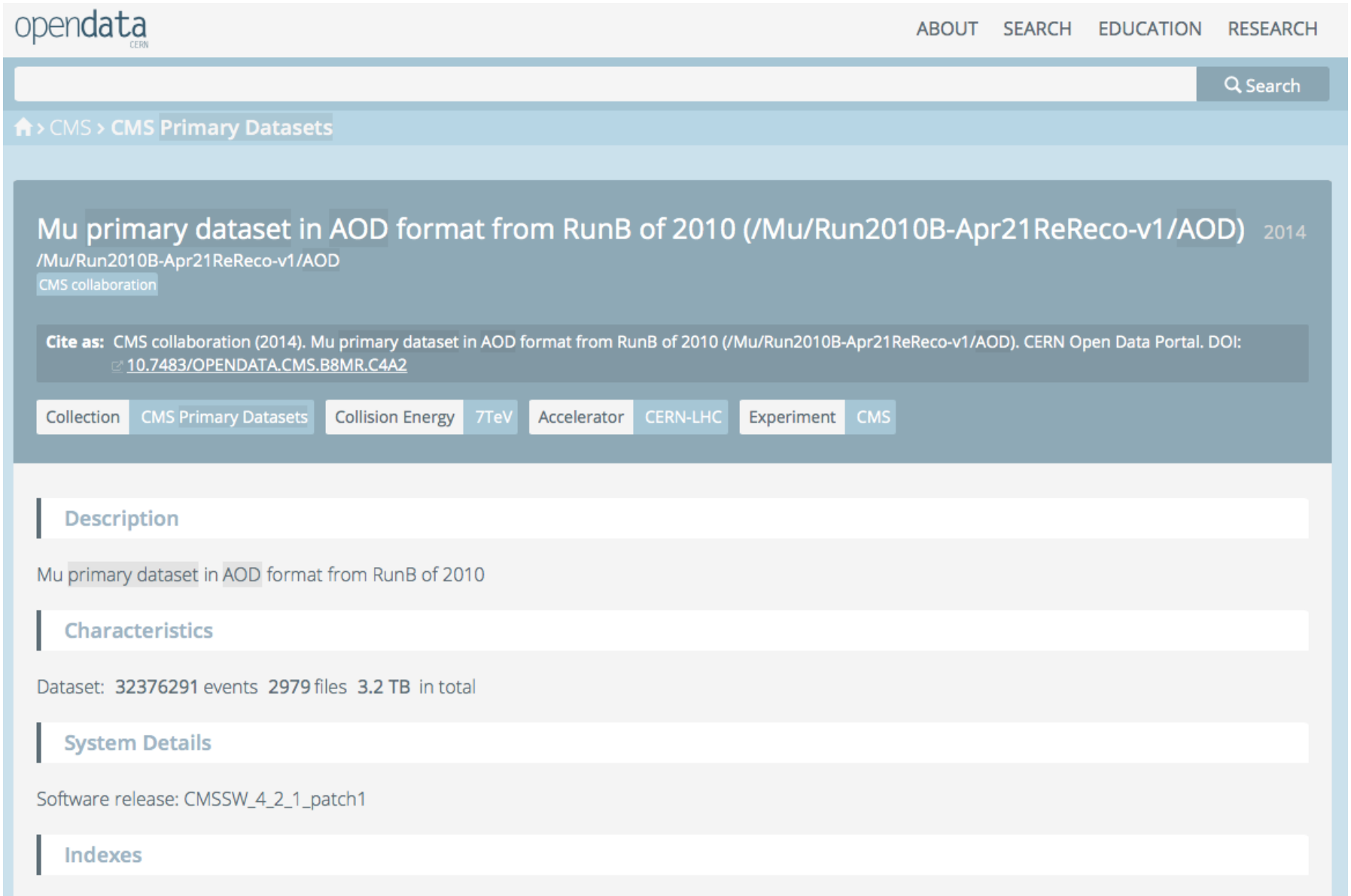
For research purposes, specific software environments and tools need to be deployed to analyse these complex primary data. In addition to the data below, you will find instructions for setting up your working environments here

Install your Virtual Machine >

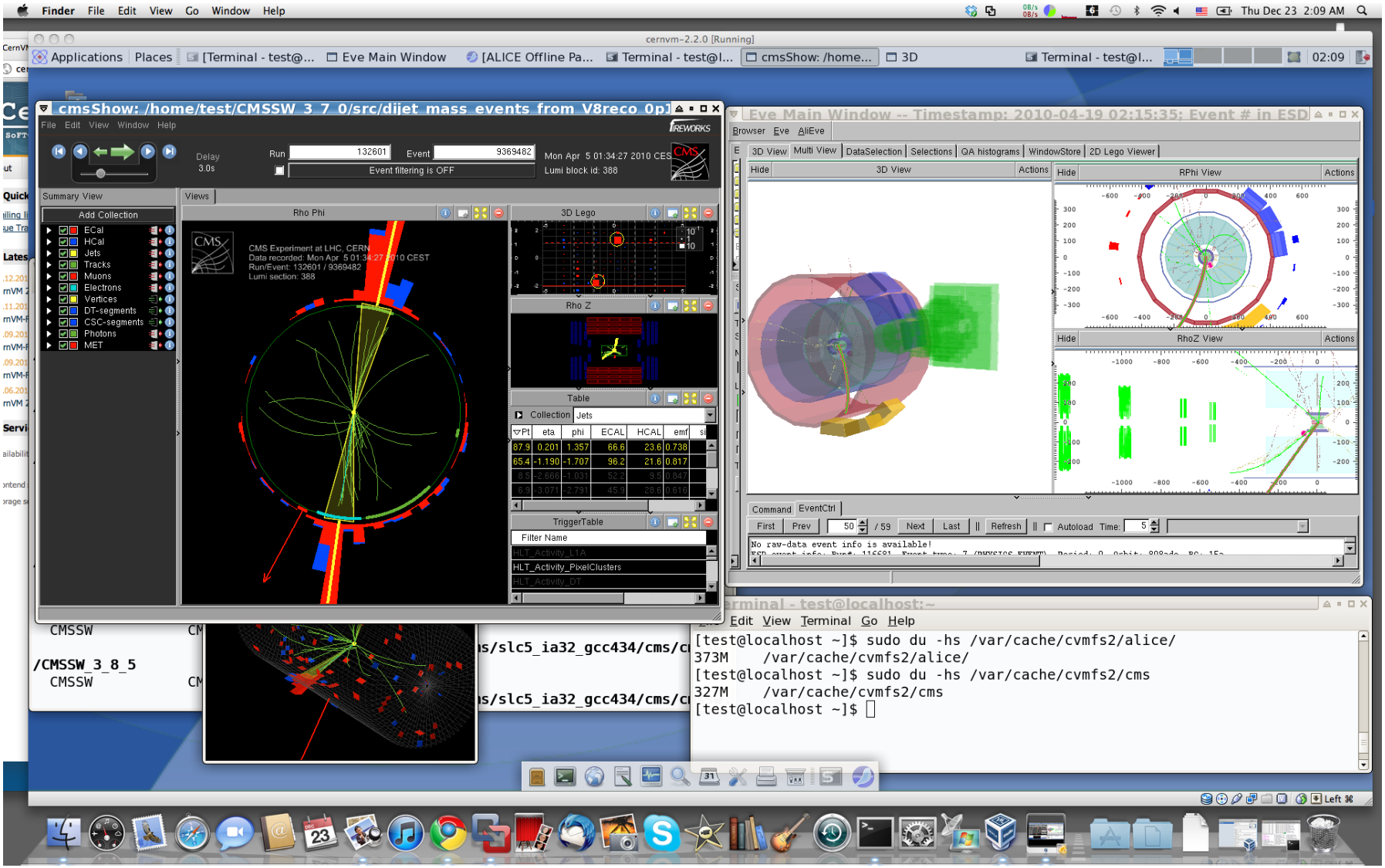


Start analysing the data >

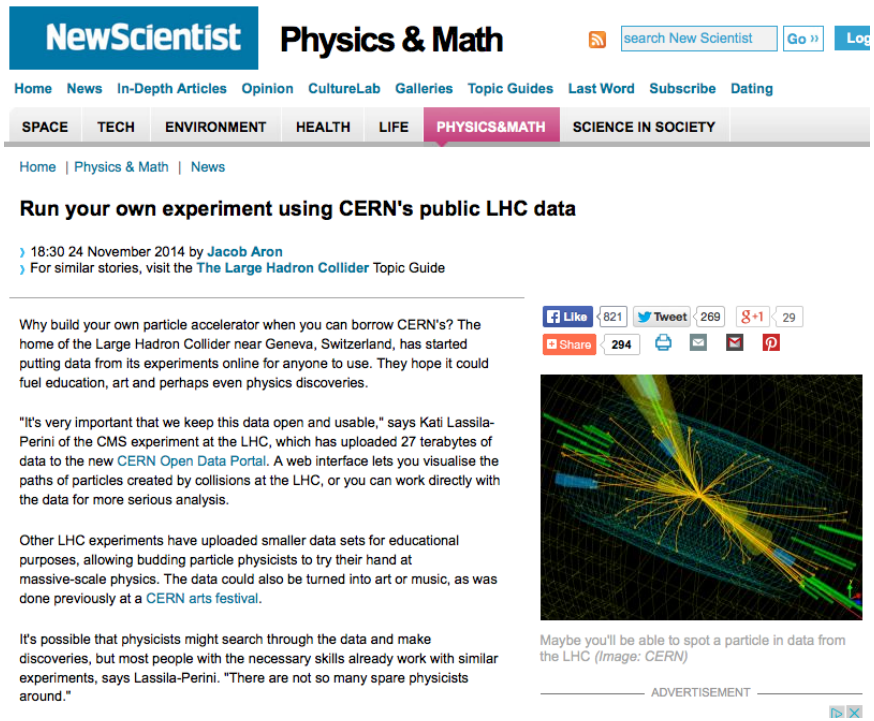
CMS Primary Datasets 2010



CernVM Virtual Machine



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.

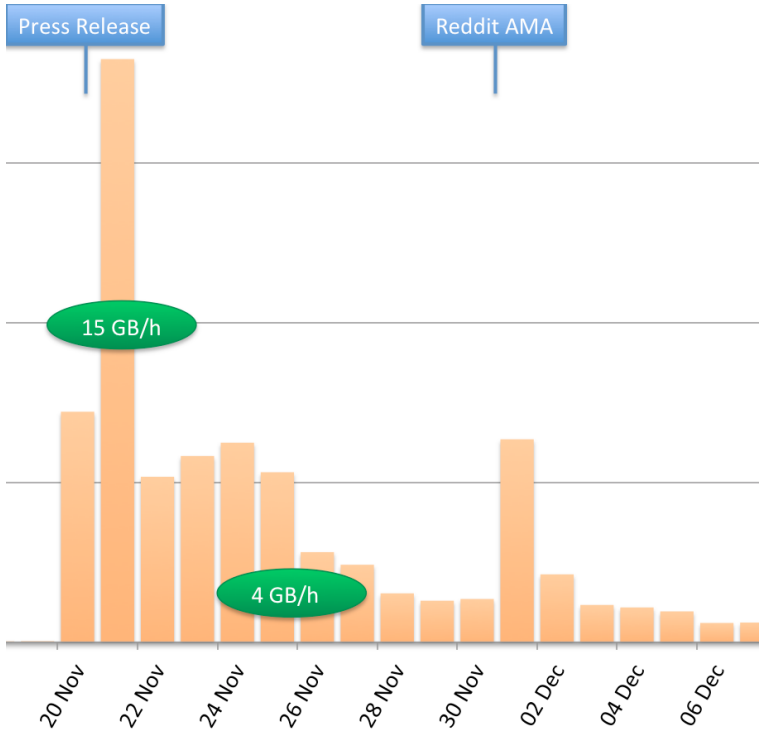
Other LHC experiments have uploaded smaller data sets for educational purposes, allowing budding particle physicists to try their hand at massive-scale physics. The data could also be turned into art or music, as was done previously at a CERN arts festival.

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



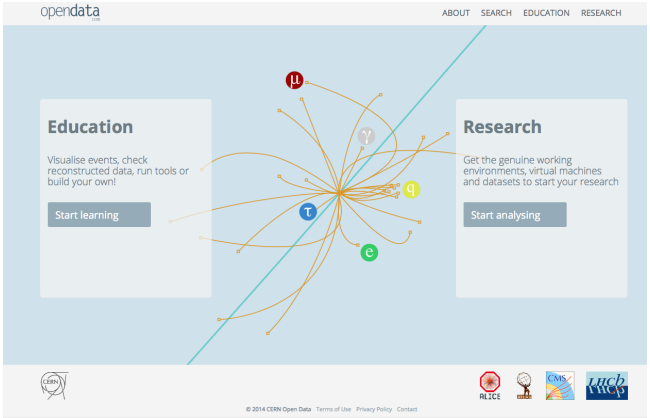
- 82,000 distinct users visited the site
- 21,000 distinct users viewed data records
- 16,000 distinct users used event display
- 3,000 distinct users used histogramming

INVENIO

7

Conclusions

CERN (Open) Data



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