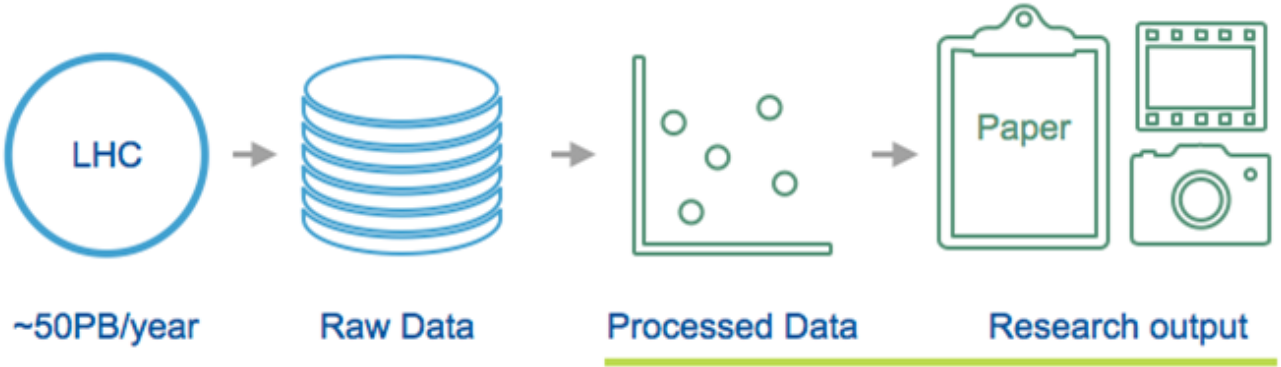
Data at CERN

Services for publishing, linking and preservation

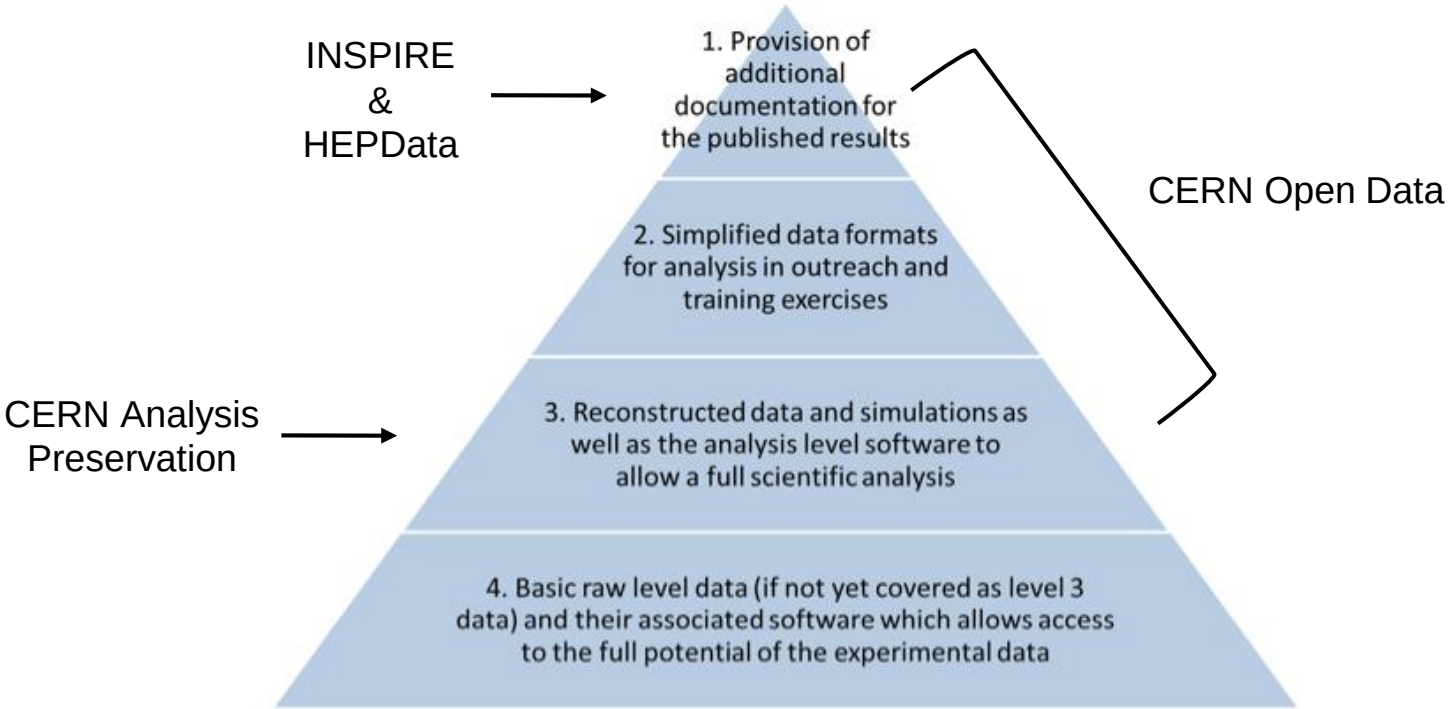
Slides prepared by Artemis Lavasa and Sünje Dallmeier-Tiessen

Presentation at the CERN-UNESCO School on Digital Libraries, Nairobi, 10th October 2018 given by Jens Vigen

Data in HEP

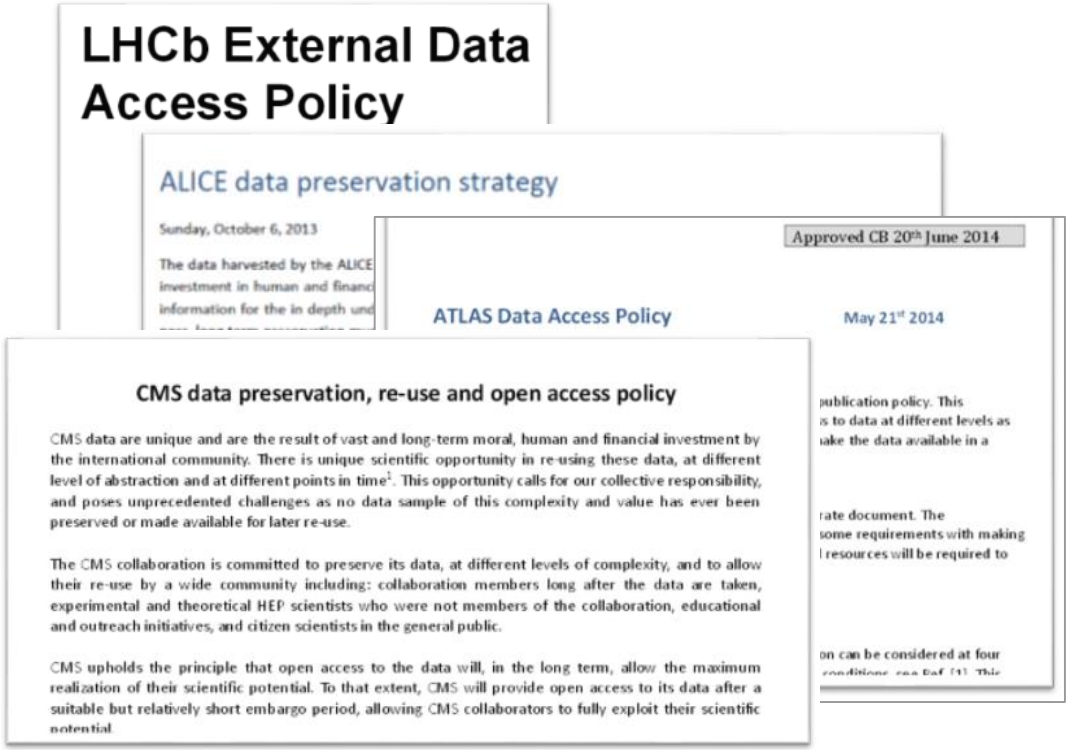

LHC
Paper
~50PB/year
Raw Data
Processed Data
Research output

Data in HEP



Herterich, P., & Dallmeier-Tiessen, S. (2016). Data Citation Services in the High-Energy Physics Community. *D-Lib Magazine*, *22*(1/2). http://doi.org/10.1045/january2016-herterich

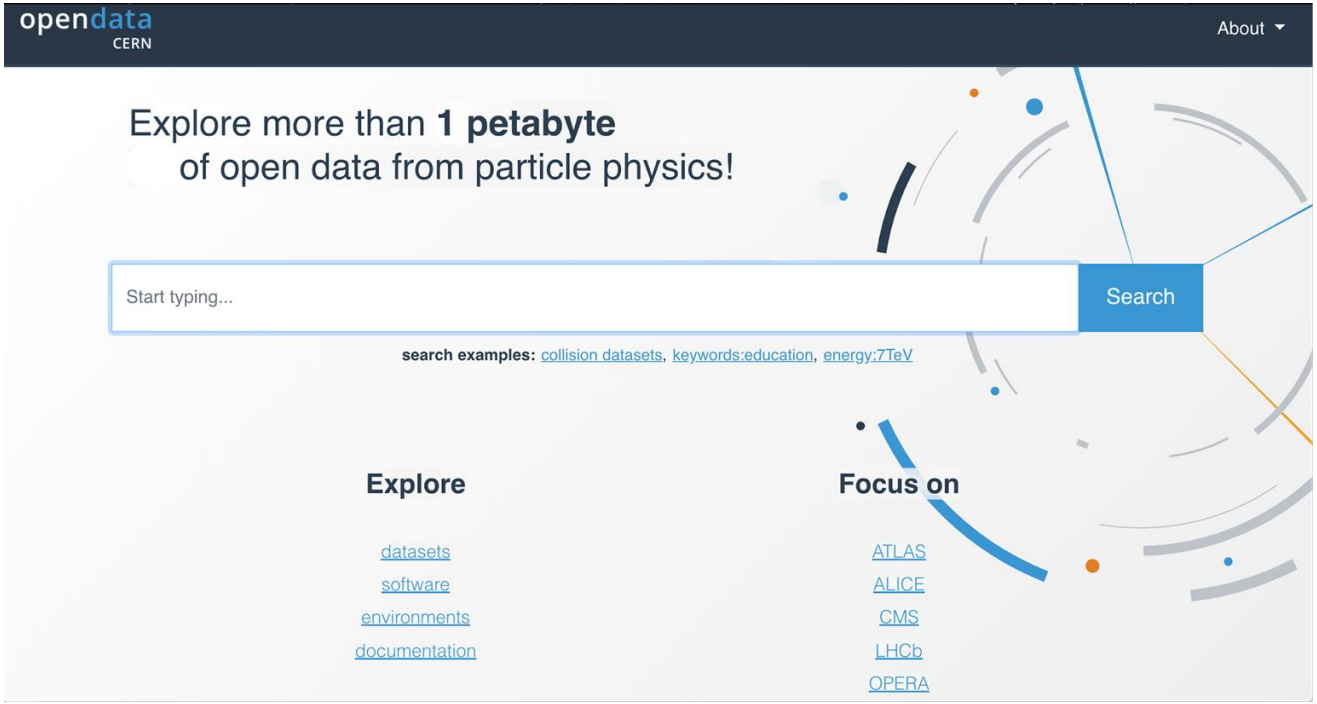
Access to HEP data: Data Policies



- ❑ Open data for three levels
- ❑ Immediate release for level 1
- ❑ Embargo periods for level 3
- ❑ CC0 – public domain dedication

Services at CERN

CERN Open Data portal (opendata.cern.ch)

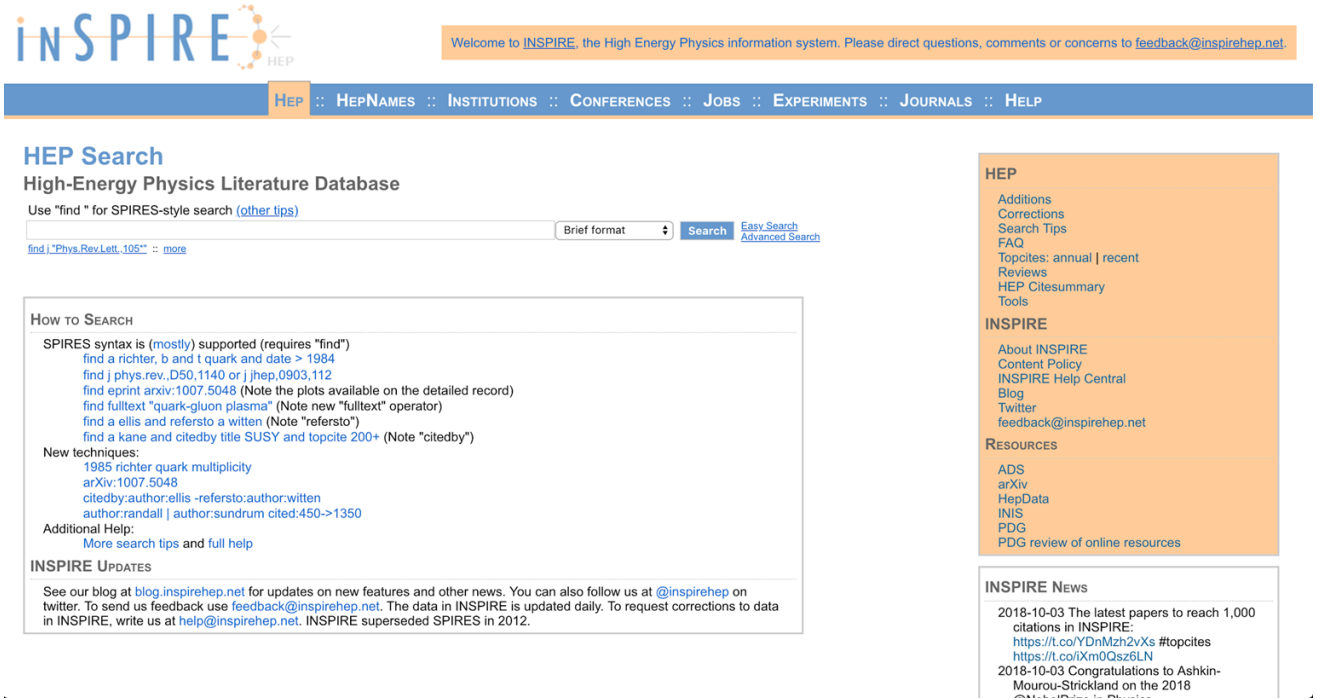


Platform for curated releases of CERN data sets, software, supplementary materials, etc., over 1PB of data

CERN Analysis Preservation

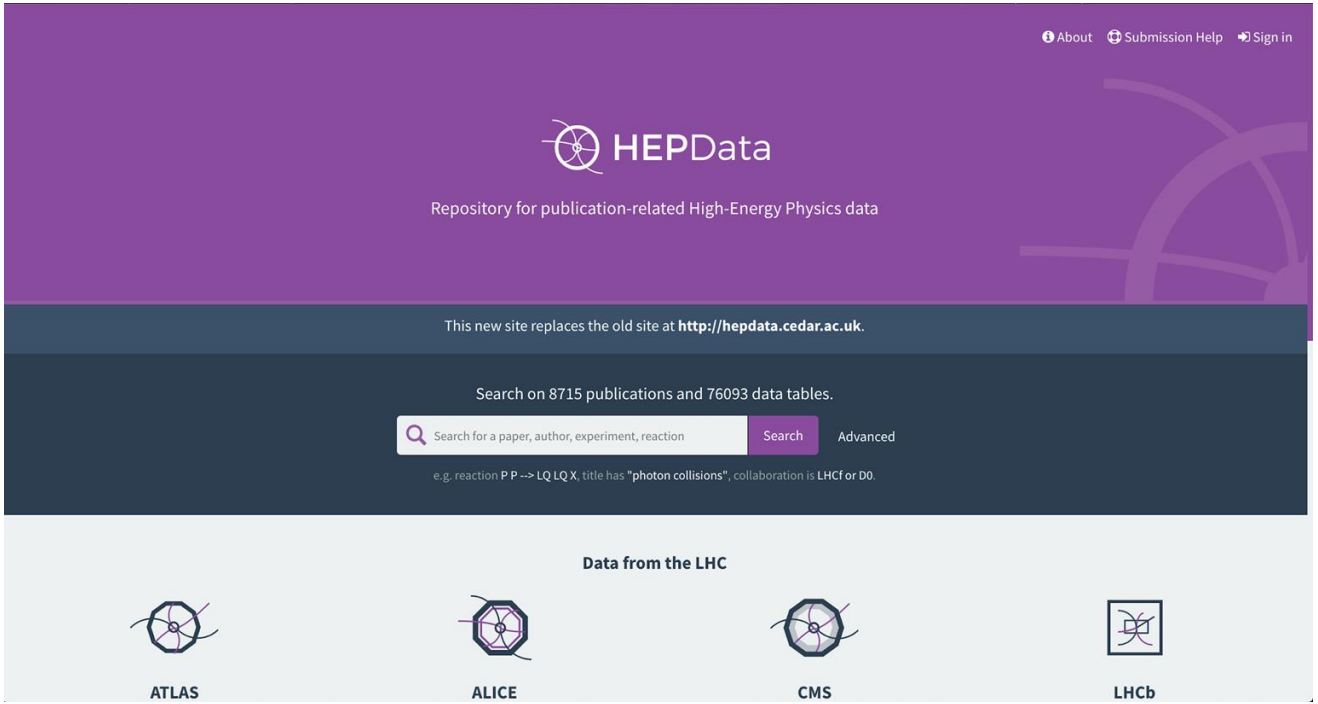
Restricted-access service, preservation of knowledge and assets of physics analyses

INSPIRE-HEP (inspirehep.net)



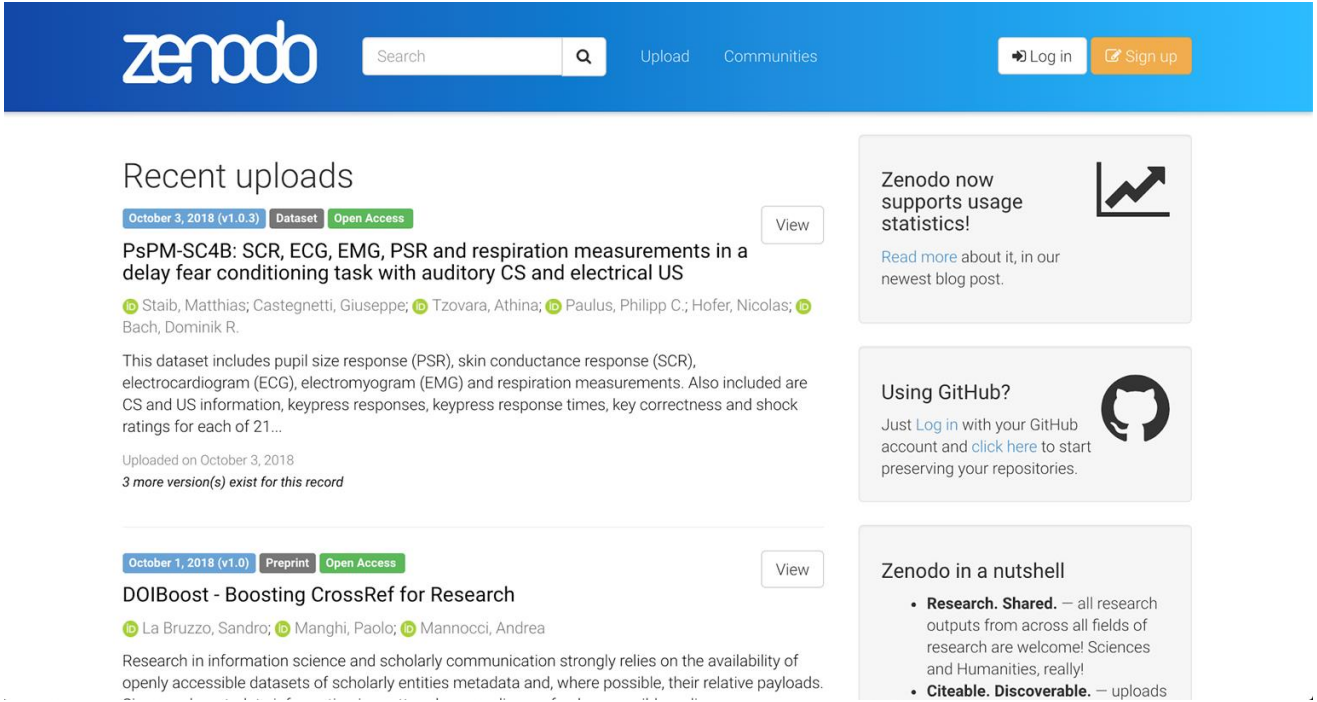
Core HEP literature aggregator, “long tail” of data

HEPData (hepdata.net)



Repository for tabular data associated with publications

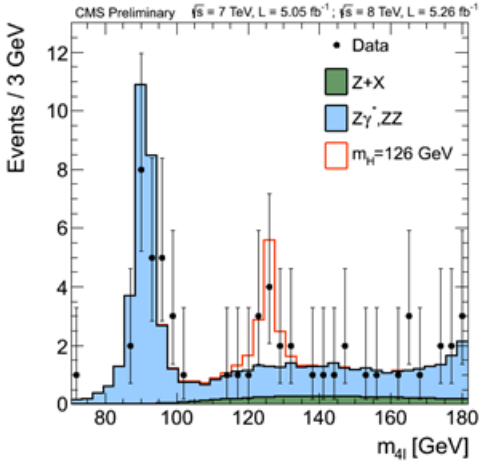
Zenodo (zenodo.org)

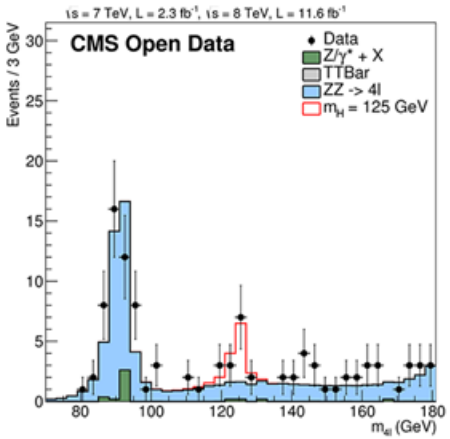


Catch-all open repository, variety of resource types, interdisciplinary

Examples of impact using CERN Open Data

- Thaler et al. used CMS open data that were released on the portal in 2014 to confirm their jet substructure model predictions
- Nur Zulaiha Jomhari et al. used CMS open data from 2011 and 2012 from the portal to run a particle physics analysis example about the Higgs decay physics using the same data formats, the same software tools and computational techniques used by CMS experimental physicists
    - On the left is is the official CMS analysis of the "Higgs-to-four-lepton" decay channel that led to the Higgs boson experimental discovery in 2012 and the plot on the right was the result of the analysis by Nur Zulaiha Jomhari et al.





https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.119.132003
https://journals.aps.org/prd/abstract/10.1103/PhysRevD.96.074003
http://opendata.cern.ch/record/5500

Features

- ➔ Developed using open-source software
- ➔ Open access licenses
- ➔ Persistent identifiers (DOIs, ORCID iDs)
- ➔ Versioning
- ➔ Machine-readable, high-quality metadata
- ➔ Provisions for big data and high complexity
- ➔ Publishing less “traditional” resources (Jupyter notebooks, Virtual Machines..)
- ➔ Enabling reuse

CERN