

# CERN IT and HEPData

Tibor Šimko

CERN

*HEPData Advisory Board Meeting, Durham, UK, 28 January 2020*

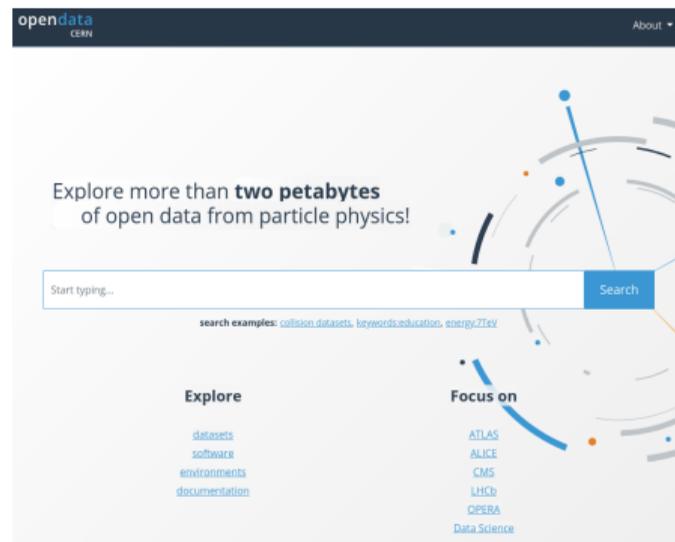
# Research data repository related activities in CERN IT

# CERN Open Data

- ▶ *event-level data for education and research*
- ▶ collision, simulated, and derived datasets
- ▶ virtual machines and container images
- ▶ software tools and analysis examples
- ▶ configuration files and documentation
- ▶ over 7K records, 800K files, 2P bytes

## HEPData

- ▶ classic content: *publication-level data*
- ▶ natural distinction and collaboration
- ▶ recent evolution: likelihoods; additional data?



<http://opendata.cern.ch>

# Zenodo

- ▶ long-tail of science; all scientific disciplines
- ▶ over 1.5M of data, software, document records
- ▶ over 70K DOIs for software
- ▶ used by individuals and self-organized communities

## HEPData

- ▶ cross-experiment communities such as Machine Learning?
- ▶ some ML data on CERN Open Data
- ▶ some ML data on Zenodo

The screenshot shows a Zenodo repository page for a dataset. The header includes the Zenodo logo, a search bar, and navigation links for 'Upload' and 'Communities'. The main title is 'CMS 2011A Simulation | Pythia 6 QCD 300-470 | pT > 375 GeV | MOD HDF5 Format'. Below the title, it shows the date 'August 8, 2019', a 'Status' badge, and '71 views' and '765 downloads'. The authors listed are Koroste, Patrick; Mastandrea, Radha; Metodiev, Eric; Nank, Prakash; and Thaler, Jesse. A detailed description follows, explaining that the dataset consists of simulated QCD jets from the CMS 2011 Open Data, processed into the MOD HDF5 format. It mentions that jets are provided at generator (part) level in the GEN files and after GEN4 detector simulation in the SIM files. The dataset includes associated GEN jets for studies involving both types of jets, and jets are selected from the hardest two anti-k<sub>R</sub> R=0.5 jets in events passing the Jet300 High Level Trigger. The jets are required to have p<sub>T</sub><sup>jet</sup> > 375 GeV, where p<sub>T</sub><sup>jet</sup> includes a jet energy correction factor (JEC), only relevant for SIM. GEN jets contain full-level particles with kinematic and POG ID information, and SIM jets contain Particle Flow Candidates (PFCs) with kinematic, POG ID, and vertex information. Additionally, jets have metadata describing their kinematics and provenance in the original CMS AOD files. There is a section for 'Additional details about the dataset', a 'Supported method for downloading, reading, and using this dataset' section mentioning the EnergyFlow Python package, and a 'For reference, the other corresponding datasets of simulated jets available on Zenodo are:' section with a list of related datasets. A table lists 10 files, each with a name, size, and download link. The right sidebar contains metadata including 'Publication date: August 8, 2019', 'DOI: 10.5281/zenodo.3541488', 'Keywords: jets, simulation, open data, the jet substructure', 'Related identifiers: arXiv:1908.05542', 'License: CC BY 4.0 International', 'Versions: Version v0 10.5281/zenodo.3541488', and 'Share' options.

zenodo Search Upload Communities Log in Register

August 8, 2019 Status Open Access

CMS 2011A Simulation | Pythia 6 QCD 300-470 | pT > 375 GeV | MOD HDF5 Format

Koroste, Patrick Mastandrea, Radha Metodiev, Eric Nank, Prakash Thaler, Jesse

Simulated QCD jets from the Simulated QCD 300-470 Dataset of the CMS 2011 Open Data reprocessed into the MOD HDF5 format. Jets are provided at generator (part) level in the GEN files and after GEN4 detector simulation in the SIM files (which also contain associated GEN jets to facilitate studies involving both types of jets). Jets are selected from the hardest two anti-k<sub>R</sub> R=0.5 jets in events passing the Jet300 High Level Trigger (only relevant for SIM) and are required to have p<sub>T</sub><sup>jet</sup> > 375 GeV, where p<sub>T</sub><sup>jet</sup> includes a jet energy correction factor (JEC), only relevant for SIM. GEN jets contain full-level particles with kinematic and POG ID information, and SIM jets contain Particle Flow Candidates (PFCs) with kinematic, POG ID, and vertex information. Additionally, jets have metadata describing their kinematics and provenance in the original CMS AOD files.

For additional details about the dataset, please see the accompanying paper, Exploring the Space of Jets with CMS Open Data. These jets were further restricted to have |y<sup>jet</sup>| < 1.9 to ensure tracking coverage and (in the case of SIM) have "medium" quality to reject fake jets.

The supported method for downloading, reading, and using this dataset is through the EnergyFlow Python package, which has additional documentation about how to read and use this and related datasets. Should any problems be encountered, please submit an issue on GitHub.

For reference, the other corresponding datasets of simulated jets available on Zenodo are:

- SM-GEN-QCD\_Jets\_170-200\_GeV
- SM-GEN-QCD\_Jets\_300-470\_GeV
- SM-GEN-QCD\_Jets\_470-600\_GeV
- SM-GEN-QCD\_Jets\_800-800\_GeV
- SM-GEN-QCD\_Jets\_800-1000\_GeV
- SM-GEN-QCD\_Jets\_1000-1400\_GeV
- SM-GEN-QCD\_Jets\_1400-1800\_GeV
- SM-GEN-QCD\_Jets\_3000-nc\_GeV

There is an associated dataset of jets recorded by the CMS detector available on Zenodo:

- CMS 2011A\_Jets\_pT > 375 GeV

Name	Size	Download
GEN300_pT375-ntGeV_0_compressed.H5	102.3 MB	Download
rd51216f1672f656a16a16d3c55464496		
GEN300_pT375-ntGeV_10_compressed.H5	102.4 MB	Download
rd51670c9a1a616a16a16d3c55464496		
GEN300_pT375-ntGeV_11_compressed.H5	102.3 MB	Download
rd517606b823017611a616d3c55464496		
GEN300_pT375-ntGeV_12_compressed.H5	102.5 MB	Download
rd514546461757176411a616d3c55464496		
GEN300_pT375-ntGeV_13_compressed.H5	102.4 MB	Download
rd5166384630713d5c497111a616d3c55464496		

Files (19 GB)

Publication date: August 8, 2019

DOI: 10.5281/zenodo.3541488

Keywords: jets, simulation, open data, the jet substructure

Related identifiers: arXiv:1908.05542

License (for files): CC BY 4.0 International

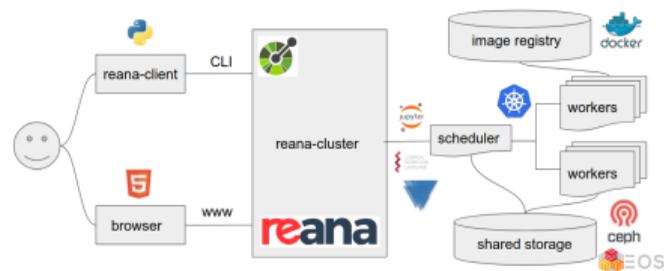
Versions: Version v0 10.5281/zenodo.3541488 Aug 8, 2019

Share: Koroste, Patrick, Mastandrea, Radha, Metodiev, Eric, Nank, Prakash, & Thaler, Jesse (2019), CMS 2011A Simulation | Pythia 6 QCD 300-470 | pT > 375 GeV | MOD HDF5 Format | Version v0 | Data set | Zenodo. <http://dx.doi.org/10.5281/zenodo.3541488>

Start typing a citation style.

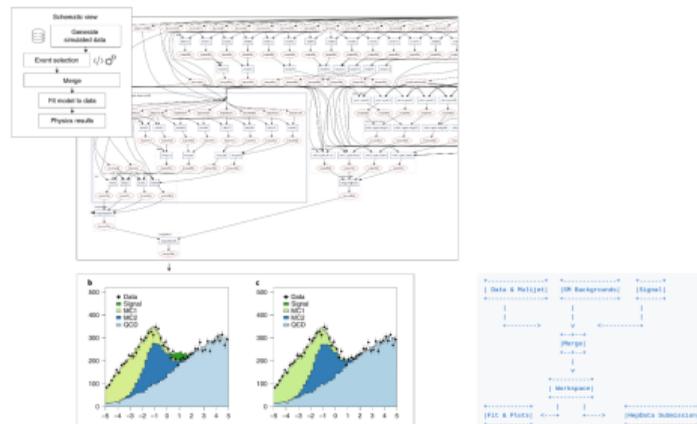
# REANA reproducible analysis platform

- ▶ run containerised analysis pipelines on remote compute clouds
- ▶ data + code + environment + workflows = reproducible science
- ▶ FAIR data reuse



## HEPData

- ▶ push workflow assets to CERN Analysis Preservation, Zenodo ... HEPData
- ▶ pull information from HEPData
- ▶ integration in automated data analysis and reuse workflows?

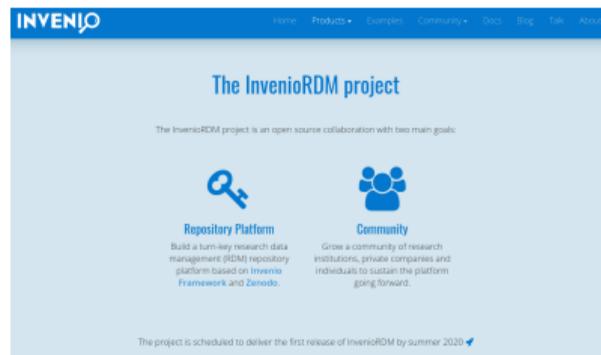


# Invenio digital repository framework

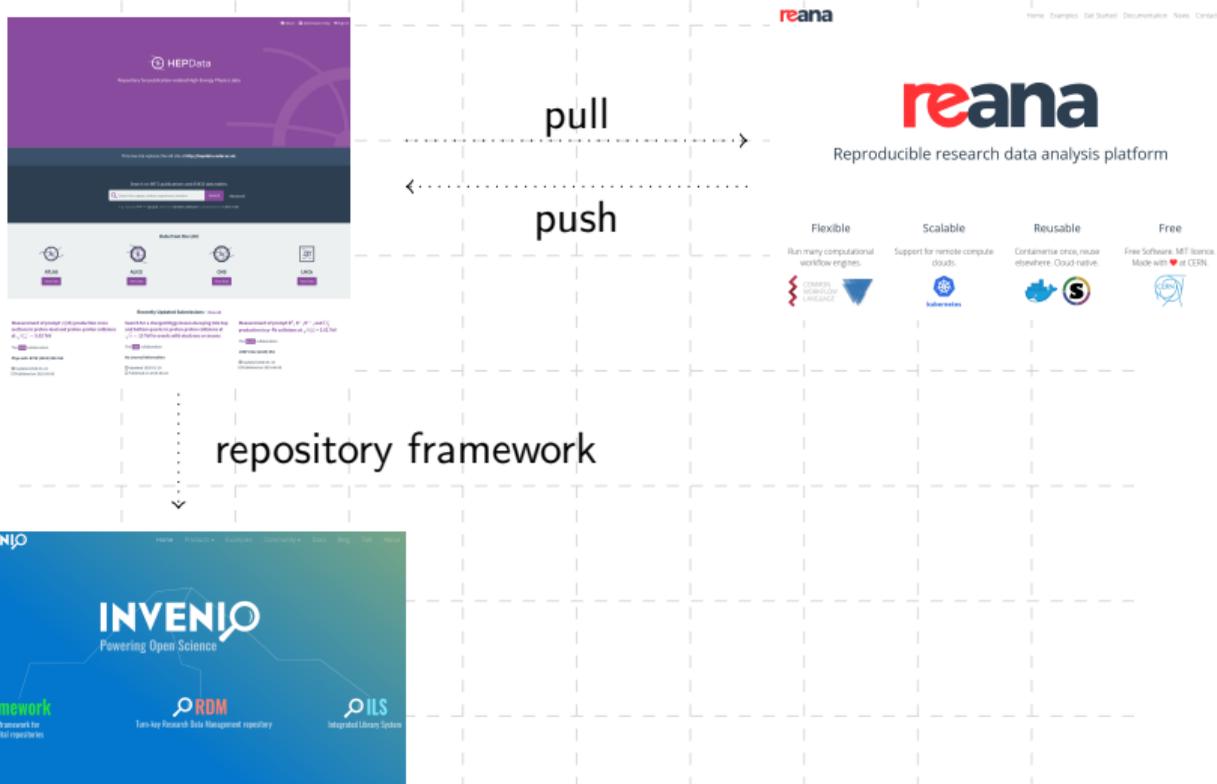
- ▶ Mature digital repository framework
- ▶ used by CERN Analysis Preservation, CERN Document Server, CERN Open Data, HEPData, INSPIRE, Zenodo . . .
- ▶ started generalised Invenio RDM project (2019-)

## HEPData

- ▶ using Invenio lightly (records module)
- ▶ interest in certain modules?
- ▶ digital repository framework needs?



# Conclusions: technology-synergies



# Conclusions: content-synergies

