HEPData

Graeme Watt (IPPP Durham)

IRIS-HEP Analysis Systems Topical Workshop

NYU Physics Department, Thursday 20th June 2019

https://hepdata.net

Email: info@hepdata.net

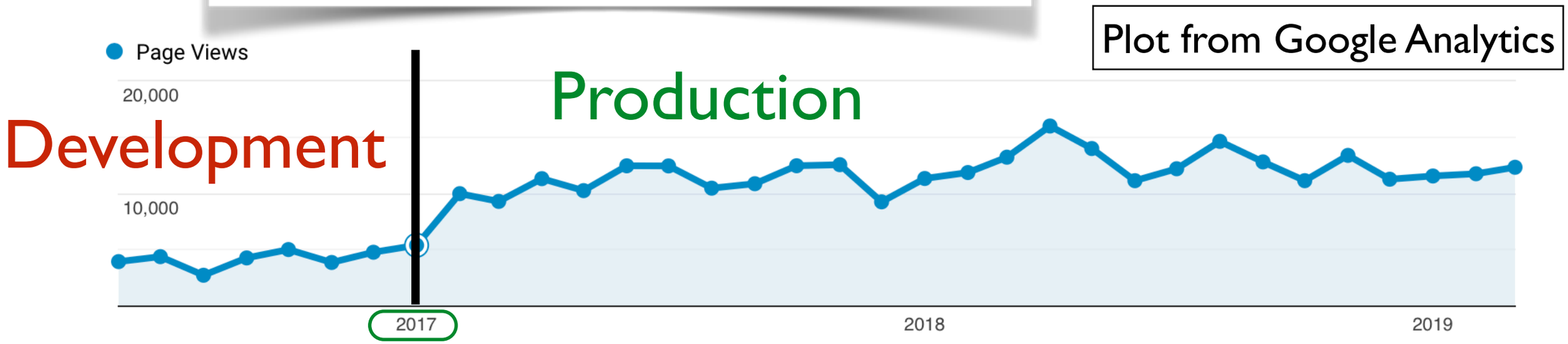
Follow @HEPData

Code: https://github.com/HEPData

What is HEPData?

- Unique *open-access* repository for scattering **data** from almost 9000 experimental **HEP** ("`hep-ex`") papers.
- Publication-related data complementary to event-level data provided through recent CERN Open Data portal.
- Traditional focus on unfolded measurements, but in recent years also include material for *recasting* LHC **searches**.
- Based in **I**nstitute for **P**article **P**hysics **P**henomenology (**IPPP**) at Durham University (UK), going back to 1970s.
- Funded by UK **S**cience & **T**echnology **F**acilities **C**ouncil (**STFC**), with current 3-year PPGP(E) grant until 09/2019. Last application **not funded**, but stop-gap of 6 months.

Transition to hepdata.net



- Software completely rewritten (2015-2016) with new hepdata.net site replacing previous hepdata.cedar.ac.uk.
- Partnership with CERN Scientific Information Service. Lead developer: Eamonn Maguire (03/2015–10/2016).
- Started from a fork of Zenodo code. Overlay on Invenio v3.
- hepdata.net hosted on CERN OpenStack infrastructure.
- External data submissions from January 2017 onwards.

CHEP 2016 paper

HEPData: a repository for high energy physics data

Eamonn Maguire[1], **Lukas Heinrich**[2] **and Graeme Watt**[3]

[1] CERN, Geneva, Switzerland
[2] Department of Physics, New York University, New York, USA
[3] IPPP, Department of Physics, Durham University, Durham, UK

E-mail: `info@hepdata.net`

Abstract. The Durham High Energy Physics Database (HEPData) has been built up over the past four decades as a unique open-access repository for scattering data from experimental particle physics papers. It comprises data points underlying several thousand publications. Over the last two years, the HEPData software has been completely rewritten using modern computing technologies as an overlay on the Invenio v3 digital library framework. The software is open source with the new site available at `https://hepdata.net` now replacing the previous site at `http://hepdata.cedar.ac.uk`. In this write-up, we describe the development of the new site and explain some of the advantages it offers over the previous platform.

1. Introduction

The Durham High Energy Physics Database (HEPData), a unique open-access repository for scattering data from experimental particle physics papers, has a long history dating back to the 1970s. It currently comprises data related to several thousand publications including those from the Large Hadron Collider (LHC). These are generally the numbers corresponding to the data points either plotted or tabulated in the publications, "Level 1" according to the DPHEP [1] classification, and HEPData is therefore complementary to the recent CERN Open Data Portal (`http://opendata.cern.ch`) which focuses on the release of data from Levels 2 and 3. The traditional focus of HEPData has been on measurements such as production cross sections and so the domain differs from the compilation of particle properties provided by the Particle Data Group (`http://www-pdg.lbl.gov`). In recent years HEPData has expanded beyond the traditional (unfolded and background-subtracted) measurements to also include data relevant for "recasting" LHC searches for physics beyond the Standard Model. The scope of HEPData is also being broadened to include data from particle decays and neutrino experiments, and potentially low-energy data relevant for tuning of the Geant4 detector simulation toolkit.

The HEPData project last underwent a major redevelopment around a decade ago [2], as part of the work of the CEDAR collaboration [3], where data was migrated from a legacy hierarchical database to a modern relational database (MySQL) and a web interface built on CGI scripts was replaced by a Java-based web interface. The old HepData site (`http://hepdata.cedar.ac.uk`) ran on a single machine hosted at the Institute for Particle Physics Phenomenology (IPPP) at Durham University. Over the last two years, a complete rewrite has once again been undertaken to use more modern computing technologies. The new site (`https://hepdata.net`) is hosted on a number of machines provided by CERN OpenStack and offers several advantages and new features compared to the old site. In this write-up, we describe the development of the new

7. Future plans

While HEPData has so far only been used for data associated with experimental particle physics papers, it could easily be used to store numerical values of theoretical predictions and related material from particle physics phenomenology papers, without any necessary changes to the software or submission workflow. There is potential to store low-energy data from nuclear, atomic, and medical physics, relevant for validation of the Geant4 (`http://geant4.cern.ch`) detector simulation toolkit, but further software development may first be needed to support keywords specific to the low-energy data and to support creation of records where the associated publications do not appear in the Inspire HEP literature database.

In future we plan to support a mixed YAML/ROOT input format where metadata is provided in YAML files (as before), but numerical values are extracted from ROOT objects and converted to the standard YAML format. HistFactory [7] is a framework used in many ATLAS studies for statistical analysis (such as determining exclusion contours). It encodes the full likelihood (including systematic uncertainties) of a measurement using semantic XML and histograms stored in ROOT files. Some preliminary work has been done to extract HEPData tables in the standard YAML format directly from a HistFactory configuration. Furthermore, work has begun on expanding the set of natively supported data types beyond a simple table to allow for richer datasets such as HistFactory configurations or simplified likelihoods [8]. The archival of such likelihood data in a lossless format could then be used by various reinterpretation packages.

8. Summary

The software underlying the Durham High Energy Physics database (HEPData) has been completely rewritten over the last two years, predominantly in the Python and JavaScript programming languages, as an overlay on the Invenio v3 digital library framework, but with a very large degree of customisation. The new site (`https://hepdata.net`) is now hosted at CERN on the OpenStack infrastructure, but still managed remotely from Durham. The transition from the old site (`http://hepdata.cedar.ac.uk`) has effectively been completed, with all data records being migrated to the new site. The new submission system has successfully been used for external data submissions from January 2017 onwards.

In conclusion, the new HEPData site provides a state-of-the-art web platform for particle physicists to make their data *Findable*, *Accessible*, *Interoperable*, and *Reusable* according to the FAIR principles (see `https://www.force11.org/group/fairgroup/fairprinciples`).

Acknowledgments

HEPData is funded by a grant from the UK Science and Technology Facilities Council. The DOI minting originates from the THOR project, funded by the European Commission under the Horizon 2020 programme. We are indebted to Mike Whalley for his dedicated 34 years of service as Database Manager for previous incarnations of the HEPData project, and for his assistance in migrating the data to the new platform. We thank Alicia Boya García, Kyle Cranmer, Sünje Dallmeier-Tiessen, Frank Krauss, Salvatore Mele, Laura Rueda, Jan Stypka and Michał Szostak for their various contributions during the redevelopment process.

References

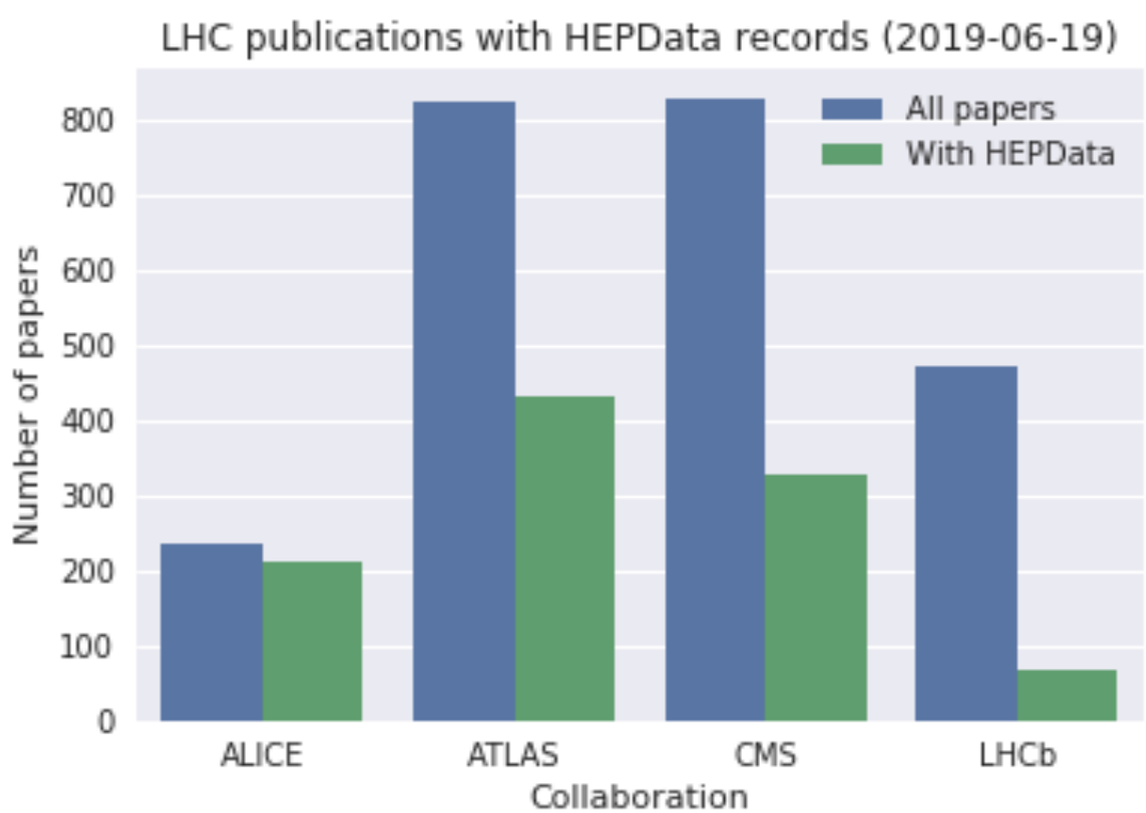
[1] Mount R *et al.* (DPHEP Study Group) 2009 (*Preprint* `0912.0255`)
[2] Buckley A and Whalley M 2010 *PoS* **ACAT2010** 067 (*Preprint* `1006.0517`)
[3] Buckley A 2007 *PoS* **ACAT2007** 050 (*Preprint* `0708.2655`)
[4] Buckley A *et al.* (Rivet) 2013 *Comput. Phys. Commun.* **184** 2803–2819 (*Preprint* `1003.0694`)
[5] Szostak M F 2015 URL `https://cds.cern.ch/record/2055193`
[6] Bonanomi M and Marcoli M 2016 URL `https://doi.org/10.5281/zenodo.197109`
[7] Cranmer K *et al.* 2012 URL `https://cds.cern.ch/record/1456844`
[8] CMS 2017 URL `https://cds.cern.ch/record/2242860`

J. Phys.: Conf. Ser. **898** 102006

[arXiv:1704.05473]

Coverage from inspirehep.net

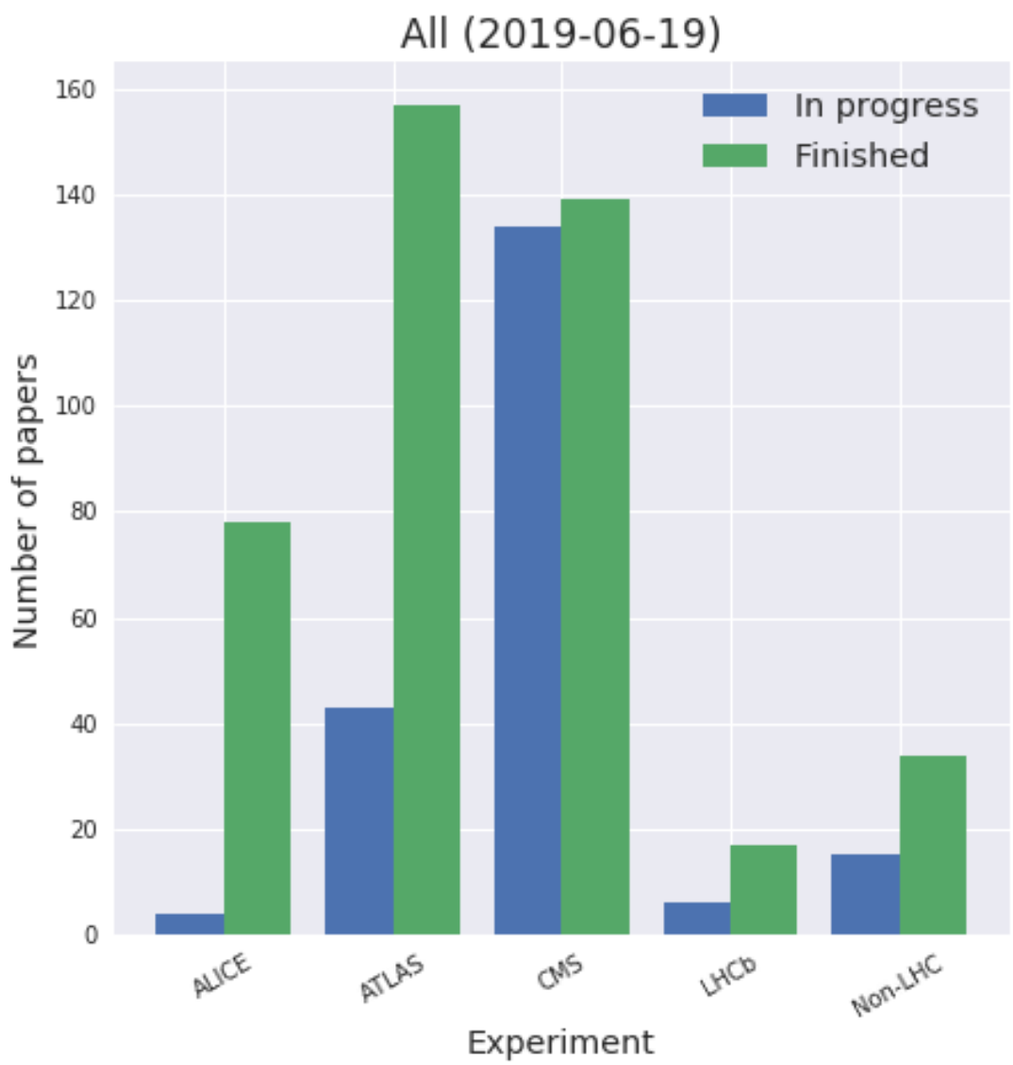
- New INSPIRE beta in development (blog) using Invenio v3.
- Legacy INSPIRE records updated to link to hepdata.net.
- Search INSPIRE for HEPData records with "035:hepdata".
- LHC publications with HEPData (Jupyter Notebook):

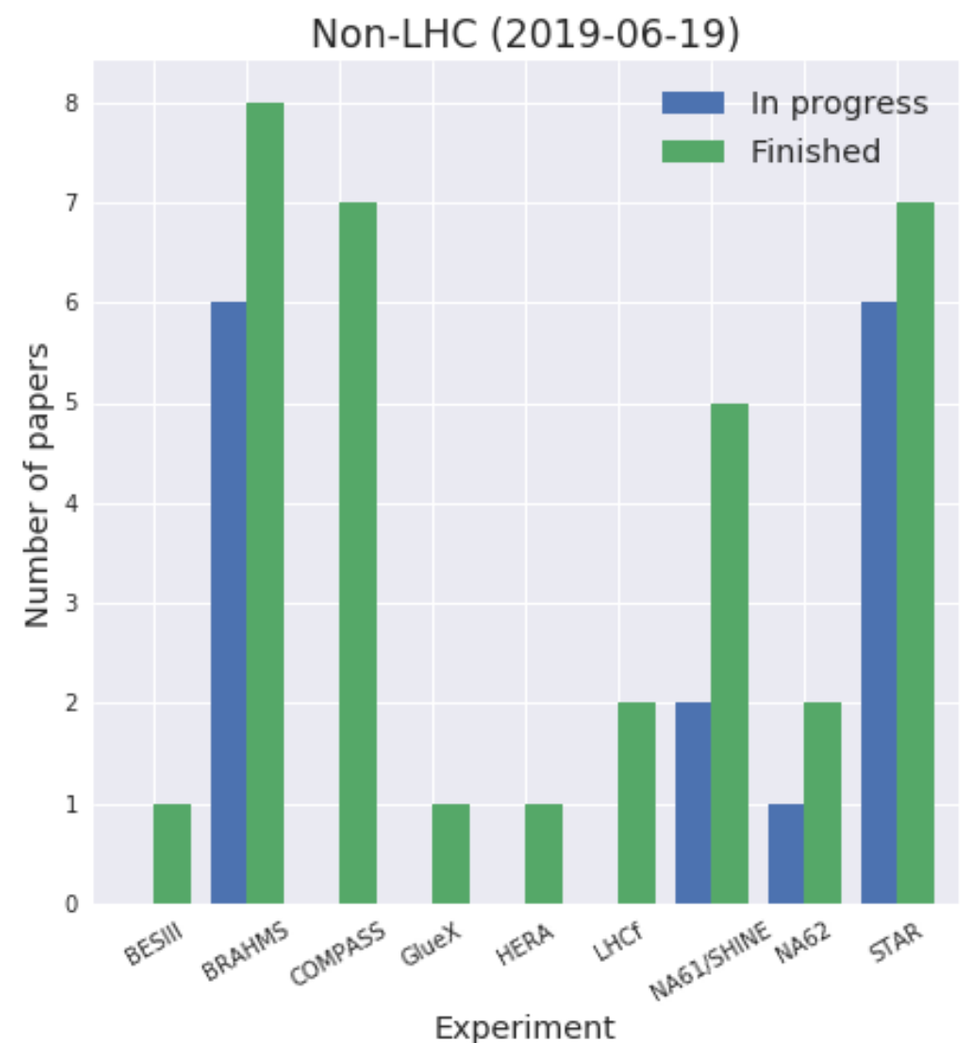


INSPIRE search query:
- `hep-ex` or `nucl-ex`
- Published in a journal
- Not conference paper

ALICE: 90%
ATLAS: 52%
CMS: 39%
LHCb: 14%

Submission system usage (01/2017-)





- Total of **425** finished submissions, comprising ALICE (78), ATLAS (157), CMS (139), LHCb (17), Non-LHC (34).
- All **7** ATLAS and **9** CMS physics groups submit to HEPData.

Links to analysis code

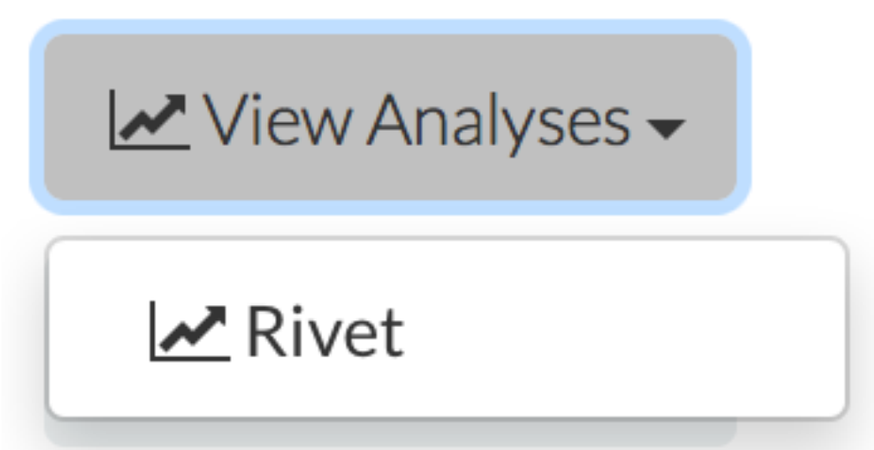
http://rivet.hepforge.org/analyses.json

- JSON file maps INSPIRE IDs to Rivet analysis names:

```
  {"422172": [“SLD_1996_S3398250”],
…,”1304688": [“ATLAS_2014_I1304688”]}
```

- Badge appears in search results and link on record:

Rivet Analysis Measurement of the $t\bar{t}$ production cross-section as a function of jet multiplicity and jet transverse momentum in 7 TeV proton-proton collisions with the ATLAS detector



- Extendable to other analysis frameworks containing publication-specific code.

pyhf likelihoods

See talks by:
Lukas Heinrich
Giordon Stark

- Recent Python implementation (pyhf) of HistFactory.
- ROOT/XML workspace replaced by plain-text pyhf JSON.
- Possible options for HEPData integration of pyhf JSON files:

1. **Host** pyhf JSON externally (cf. Rivet) linked from HEPData.
2. **Attach** pyhf JSON as auxiliary files to HEPData records.
3. **Convert** pyhf JSON file to usual HEPData YAML format.
4. **Native** support of pyhf JSON as alternative data format:
   - ✦ Replace usual HEPData YAML data files with pyhf JSON.
   - ✦ Validate against JSON schema distributed with pyhf.
   - ✦ Develop appropriate *visualisation* and *conversion* tools.