Graeme Watt (IPPP Durham)
(Re)interpreting the results of new physics searches at the LHC
Imperial College London, Tuesday 2nd April 2019

https://hepdata.net

Email: info@hepdata.net

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 Institute for Particle Physics Phenomenology (IPPP) at Durham University (UK), going back to 1970s.
- Funded by UK Science & Technology Facilities Council (STFC) until 09/2019. Pending application from 10/2019 for two positions of Project Manager and Software Engineer.
- Software completely rewritten (2015-2016) with new hepdata.net site replacing previous hepdata.cedar.ac.uk.
- 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.
HEPData: a repository for high energy physics data

Eamonn Maguire, Lukas Heinrich and Graeme Watt
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](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 provided by the Particle Data Group (http://pdg.lbl.gov). In recent years HEPData has expanded beyond the traditional (unfolded and background-subtracted) measurements to also include data relevant for “recoiling” LHC searches for physics beyond the Standard Model. The scope of HEPData is also being broadened to include data from particle decay and neutrino experiments, and potentially low-energy data relevant for tuning 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 site.

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 (each as determining exclusion contours). It encodes the full likelihood (including systematic uncertainties) of a measurement using semantic YAML 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 OpenStack, and is 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 been successfully 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://w3c.forcefield.org/group/fairgroup/ fairprinciples]).

Acknowledgments

HEPData is funded by a grant from the UK Science and Technology Facilities Council. The DO1 mining 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 Boyd García, Kyle Cranmer, Sinije Dallmeier-Tiessen, Frank Krauss, Salvatore Mele, Laura Rueda, Jan Stepyn and Michal Szostak for their various contributions during the redevelopment process.

References

[7] Cranmer K et al. 2012 URL [https://dx.doi.org/10.48554/1465944]
[8] CMS 2017 URL [https://dx.doi.org/10.48550/2228980]
Data output formats

- **JSON**: JavaScript Object Notation.
- **CSV**: comma-separated values.
- **YODA**: for inclusion in a Rivet analysis.
- **ROOT**: binary `.root` file.

**YAML**: native HEPData format.

submission.yaml

+ YAML data files for each table

https://hepdata.net/record/ins1422615

- Each table in a directory.
- **TGraphAsymmErrors** for each dependent variable.
- If finite bin width, also separate **TH1F** for value and each error.
- Multidimensional (**TH2F**, **TH3F**).
DOIs and versioning

- DOIs minted for whole record and each table via DataCite.
“Watch List” and “Ask a Question”

- Feedback from kick-off meeting in 2016 implemented.
- Both features require login.

**Watch List**

Search for new resonances in events with one lepton and missing transverse momentum in $pp$ collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector


Last updated: 2016-11-15T16:16:07

**Ask a Question**

Your question will be emailed to those involved with the submission.

Send question on this data...
Publication-driven search

- Powerful search and faceting provided by Elasticsearch.
- Or find via DataCite Search or Google Dataset Search.

Append `format=json` to URL to return search results in JSON format.

https://hepdata.net/search/?q="dark+matter"&collaboration=ATLAS
Links 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):

![Graph showing LHC publications with HEPData records (2019-03-27)]

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

**Results:**
- ALICE: 90%
- ATLAS: 53%
- CMS: 39%
- LHCb: 14%
Coverage of LHC publications by year
Submission system on hepdata.net

https://hepdata.net/submission

- Also a Sandbox for any user to test uploads without special privileges.
Submission system usage (01/2017-)

- Total of 375 finished submissions:
  - ALICE (71)
  - ATLAS (149)
  - CMS (122)
  - LHCb (13)
  - Non-LHC (20)

- 58/375 with version ≥ 2

- 1086 users created an account (as of 29/03/2019).
Submissions by ATLAS/CMS physics groups

- All 7 ATLAS and 9 CMS physics groups submit to HEPData.
- New activity by CMS **EXO** and **SUS** physics analysis groups.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>In progress</td>
<td>In progress</td>
</tr>
<tr>
<td>Finished</td>
<td>Finished</td>
</tr>
</tbody>
</table>

### ATLAS Physics Group

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPHY</td>
<td></td>
</tr>
<tr>
<td>EXOT</td>
<td></td>
</tr>
<tr>
<td>HION</td>
<td></td>
</tr>
<tr>
<td>HIGG</td>
<td></td>
</tr>
<tr>
<td>STDM</td>
<td></td>
</tr>
<tr>
<td>SUSY</td>
<td></td>
</tr>
<tr>
<td>TOPQ</td>
<td></td>
</tr>
</tbody>
</table>

### CMS Physics Analysis Group

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2G</td>
<td></td>
</tr>
<tr>
<td>BPH</td>
<td></td>
</tr>
<tr>
<td>EXO</td>
<td></td>
</tr>
<tr>
<td>F5Q</td>
<td></td>
</tr>
<tr>
<td>HIN</td>
<td></td>
</tr>
<tr>
<td>HIG</td>
<td></td>
</tr>
<tr>
<td>SMP</td>
<td></td>
</tr>
<tr>
<td>SUS</td>
<td></td>
</tr>
<tr>
<td>TOP</td>
<td></td>
</tr>
</tbody>
</table>
Submission documentation

- Many code improvements to smooth submission process: propagate error message from exceptions, EOS fixes.

- New [hepdata-submission.readthedocs.io](https://hepdata-submission.readthedocs.io) web site. Includes example scripts (simple, complicated) and offline validation script.

- New [hepdata_lib](https://github.com/HEPData/hepdata_lib) by Clemens Lange and Andreas Albert. Library to read in text/ROOT and write HEPData YAML. [https://github.com/HEPData/hepdata_lib](https://github.com/HEPData/hepdata_lib)

**hepdata_lib**

**DOI**: 10.5281/zenodo.1929718  **pypi package**: 0.2.7  **build**: passing  **coverage**: 88%  **docs**: passing  **637.2MB**: 37 layers

Library for getting your data into HEPData

[Launch](#)  [Binder](#)  [Open in SWAN](#)
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:

  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.
HEPData / Rivet compatibility

- **HEPData** exports data in **YODA** format for **Rivet**.
- Observed divergence in recent years between Rivet reference data and HEPData export.
- **Python scripts** written to assess compatibility.
- **Rivet 2.6.1**: 73 of 378 (19%) were compatible.
- Scripts **rivet-diffhepdata** and **rivet-diffhepdata-all** now included in official Rivet distribution.
- **Rivet 2.7.0**: 149 of 404 (37%) were compatible.
Covariance matrices in HEPData

Correlation/covariance matrices can be encoded in a format with two independent variables (giving the bins) and one dependent variable (giving the covariance/correlation), e.g.

```json
independent_variables:
- header: {name: PTjet, units: GeV}
  values:
  - {low: 25, high: 45}
  - {low: 45, high: 65}
  - {low: 45, high: 65}
  ...
- header: {name: PTjet, units: GeV}
  values:
  - {low: 25, high: 45}
  - {low: 25, high: 45}
  - {low: 45, high: 65}
  ...
dependent_variables:
- header: {name: Correlation}
  values:
  - {value: 1.0000}
  - {value: 0.8727}
  - {value: 1.0000}
  ...
```

- Export to **JSON**, **CSV**, **ROOT** (**TGraph2DErrors**, **TH2F**), **YODA** (**Scatter3D**).
Error breakdown in HEPData

• One *independent variable* and one *dependent variable*:

```
independent_variables:
- header: {name: PT(JET), units: GEV}
  values:
    - {low: 100.0, high: 116.0}
...

dependent_variables:
- header: {name: D2(SIG)/DPT(JET)/DYRAP, units: PB*GEV**-1}
  values:
    - value: 4230.0
      errors:
        - {label: stat, symerror: 0.89%}
        - {label: 'sys,jes0', symerror: 3.0%}
        - {label: 'sys,jes1', asymerror: {plus: 0.4%, minus: -0.5%}}
...
```

https://www.hepdata.net/record/ins1325553

• Export to JSON, CSV, ROOT (TGraphAsymmErrors, TH1F), YODA (Scatter2D).
Error breakdown in YODA

• Work with Louie Corpe (talk). New converter release deployed on 26th November 2018.
Summary

- Software completely rewritten (2015-2016) with new `hepdata.net` site replacing previous `hepdata.cedar.ac.uk`.
- Continuous improvements to **HEPData** operations, software, and documentation since original transition in January 2017.
- Increasing adoption by LHC physics groups, including BSM.
- Recent efforts to better sync reference data included in **Rivet** analyses and to export error breakdown in **YODA** format.
- Pending **STFC** grant application to continue HEPData support.

**Code:** [https://github.com/HEPData](https://github.com/HEPData)

Email: info@hepdata.net

Twitter: @HEPData
HistFactory

fundamentally a (quite flexible) p.d.f template to build **statistical models**
from binned distributions and data.

~all binned, folded, frequentist statistical models in ATLAS are expressed
using HistFactory template

- Based on simple **ROOT** histograms organised in an **XML** file.
**pyhf data format**

- Recent Python implementation (pyhf) using a **JSON** schema.

---

**HistFactory-JSON Example**

```json
{
  "channels": [
    {
      "name": "singlechannel",
      "samples": [
        {
          "name": "signal",
          "data": [5.0, 6.0],
          "modifiers": [
            {"name": "mu", "type": "normfactor", "data": null}
          ]
        },
        {
          "name": "background",
          "data": [50.0, 65.0],
          "modifiers": [
            {"name": "uncorr_bkguncrt", "type": "shapesys", "data": [5.0, 3.0]}
          ]
        }
      ]
    }
  ]
}
```

---

- **Idea:** replace usual **HEPData YAML** data files with **HistFactory JSON** data files.
- **Validate w/ schema.**
- **Need to develop visualisation and conversion tools.**

---

NEW YORK UNIVERSITY

G. Watt