HEPData status

Graeme Watt (Project Manager)

HEPData advisory board meeting

IPPP Durham, 28th January 2020

https://hepdata.net

Email: info@hepdata.net

Follow @HEPData

Code: https://github.com/HEPData

HEPData advisory board

Michael Spannowsky (IPPP Durham, PI)
Jeppe Andersen (IPPP Durham, Co-I)
Daniel Maître (IPPP Durham, Co-I)
Alexander Kohls (CERN, SIS)
Stella Christodoulaki (CERN, INSPIRE)
Tibor Šimko (CERN, IT)
William Barter (Imperial, LHCb)
Henning Flaecher (Bristol, CMS)
Bill Murray (RAL/Warwick, ATLAS)
Enrico Scomparin (Torino, ALICE)
Matthew Wing (UCL, non-LHC experiments)
Andy Buckley (Glasgow, Rivet)
Kyle Cranmer (NYU, generic users)
Peter Skands (Monash, MCPLOTS)

Meetings:
24th November 2017
(25th April 2016)
16th October 2015
12th November 2014

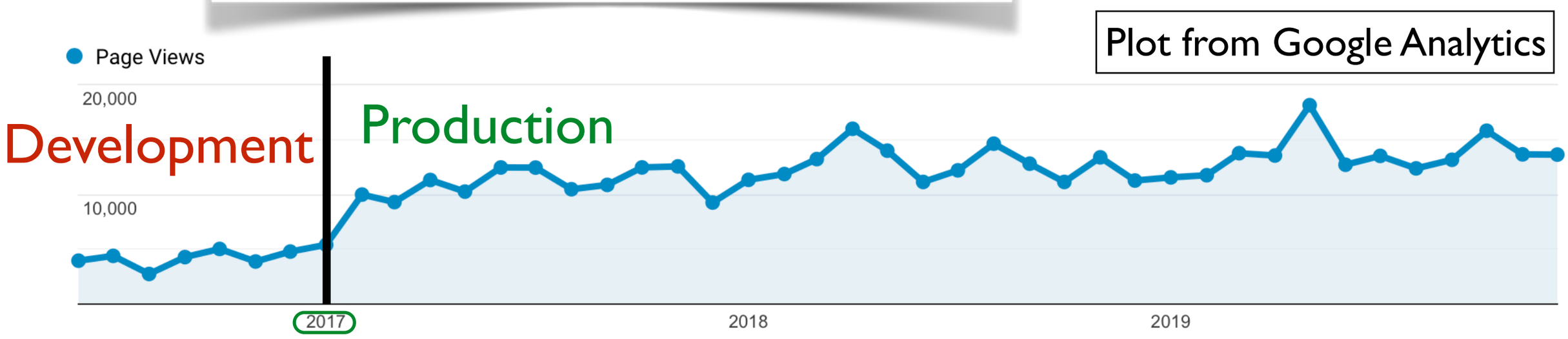
New members since 2017

Previous members in 2017
- Keith Ellis (IPPP Durham, PI)
- Salvatore Mele (CERN)
- Sünje Dallmeier-Tiessen (CERN)
- Ulrik Egede (LHCb)

What is HEPData?

- Unique *open-access* repository for high-level **data** from about 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.
- Transition in 2017 to hepdata.net site, hosted at CERN.

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.

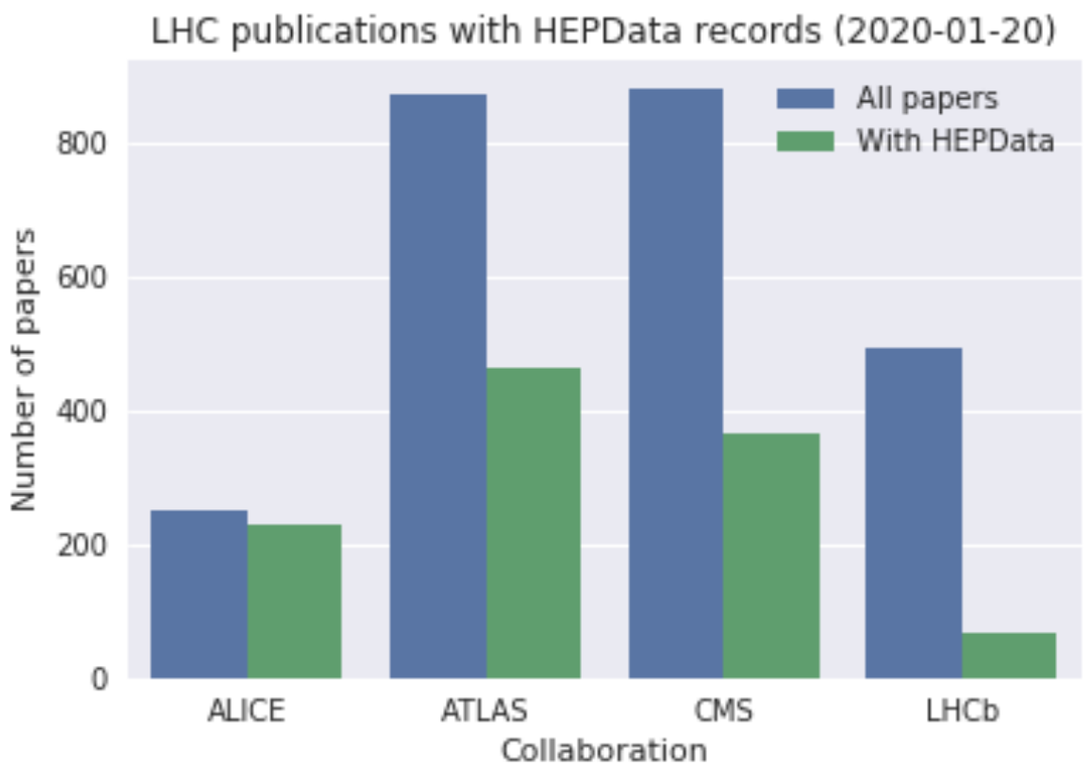
References

[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

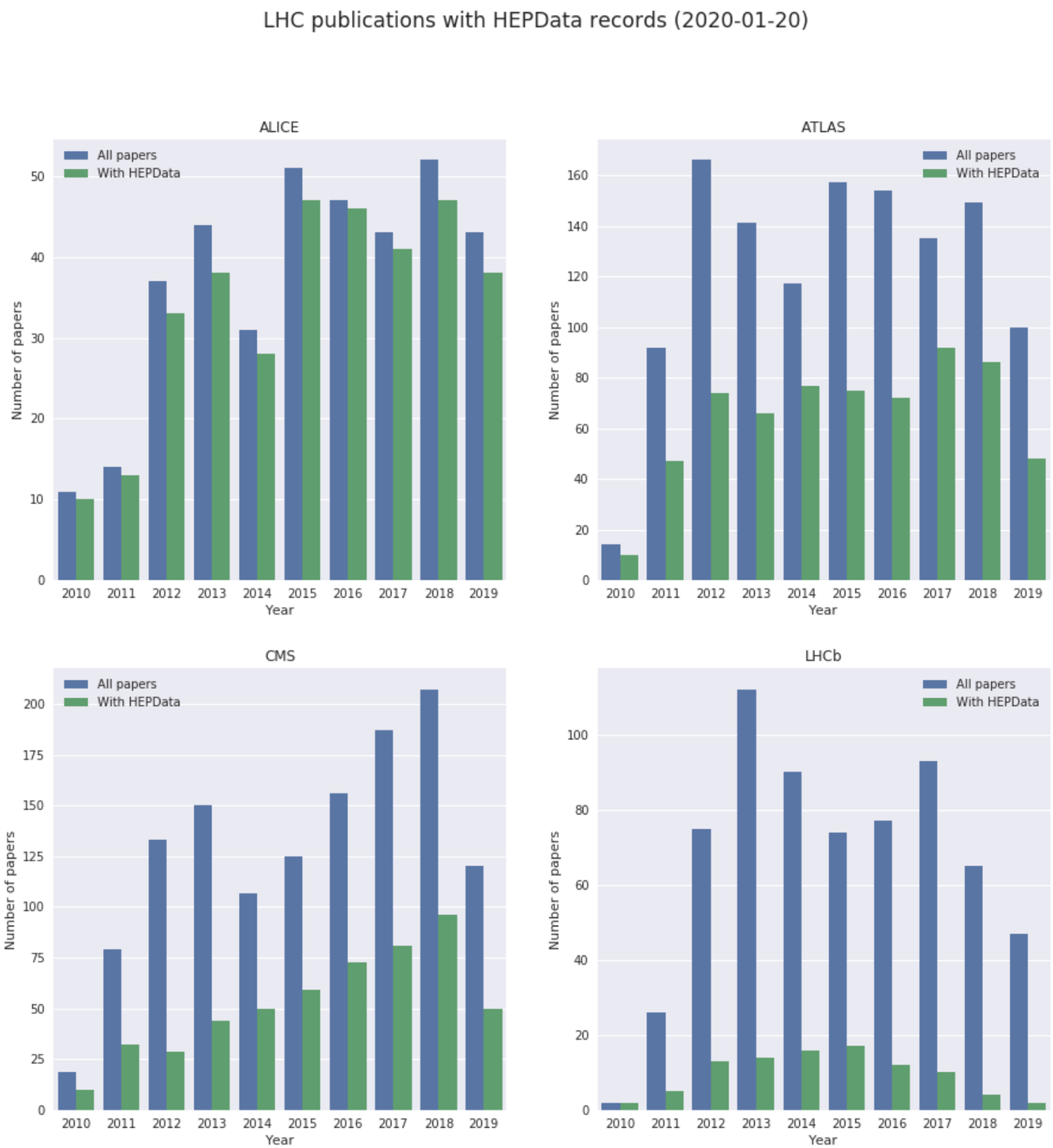
- New INSPIRE beta in development (blog) using Invenio v3.
- Legacy INSPIRE records link to hepdata.net records.
- Search INSPIRE for HEPData records with "035:hepdata".
- LHC publications with HEPData (Jupyter Notebook):



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

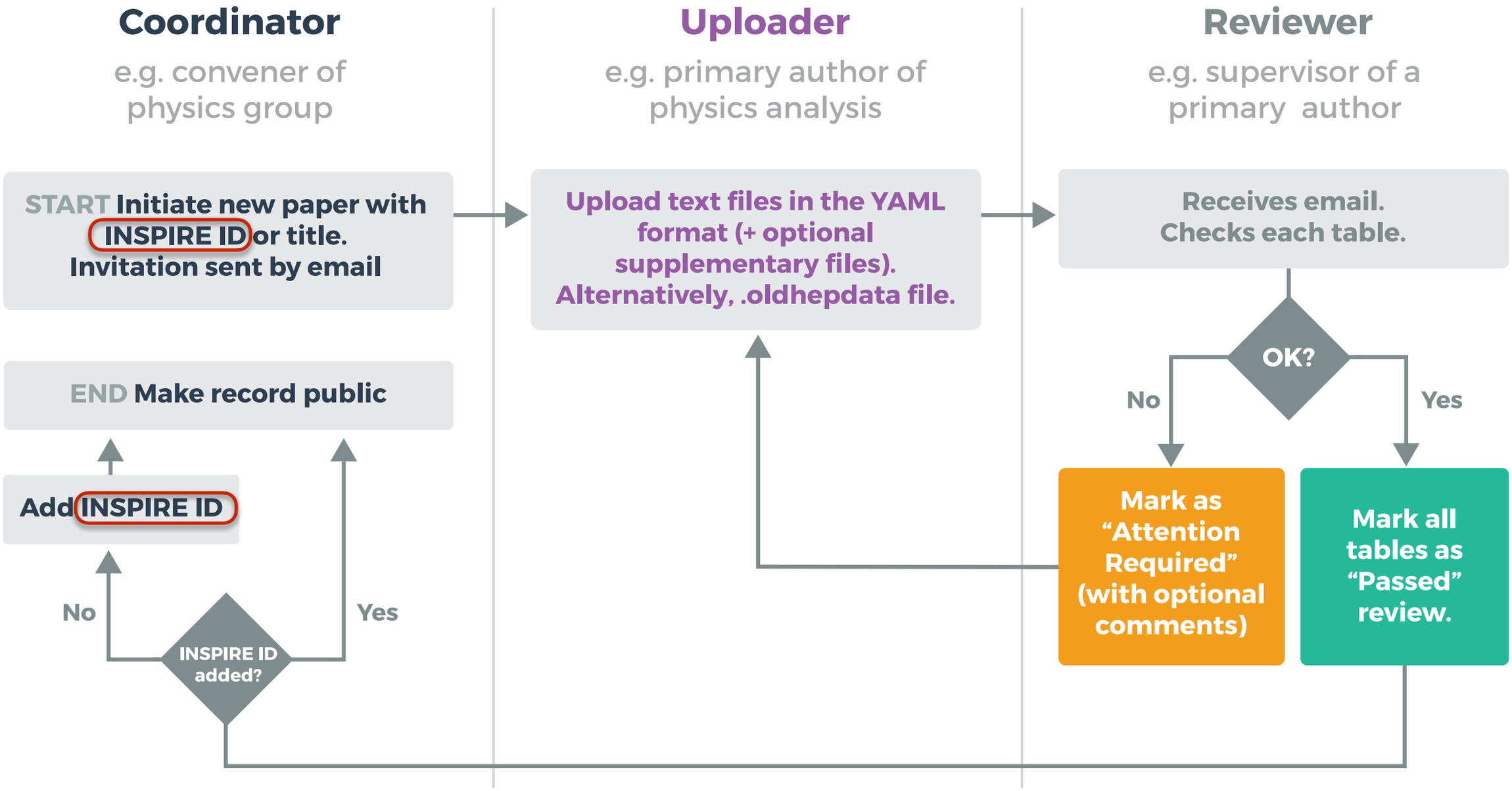
ALICE: **90%** (88% in 2017)
ATLAS: **53%** (50% in 2017)
CMS: **41%** (35% in 2017)
LHCb: **14%** (15% in 2017)

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.

Data from STEREO [arXiv:1912.06582]

Improved Sterile Neutrino Constr

Not Secure | inspirehep.net/record/1770821

INSPIRE HEP

Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

Information | References (60) | Citations (2) | Files | Plots | Data

Improved Sterile Neutrino Constraints from the STEREO Experiment with 179 Days of Reactor-On Data

STEREO Collaboration (Helena Almazán Molina (Heidelberg, Max Planck Inst.) *et al.*) Show all 27 authors

Dec 13, 2019 - 29 pages

e-Print: arXiv:1912.06582 [hep-ex] | PDF
Experiment: STEREO

Abstract (arXiv)
The STEREO experiment is a very short baseline reactor antineutrino experiment. It is designed to test the hypothesis of light sterile neutrinos being the cause of a deficit of the observed antineutrino interaction rate at short baselines with respect to the predicted rate, known as the Reactor Antineutrino Anomaly. The STEREO experiment measures the antineutrino energy spectrum in six identical detector cells covering baselines between 9 and 11 m from the compact core of the ILL research reactor. In this article, results from 179 days of reactor turned on and 235 days of reactor turned off are reported in unprecedented detail. The current results include improvements in the description of the optical model of the detector, the gamma-cascade after neutron captures by gadolinium, the treatment of backgrounds, and the statistical method of the oscillation analysis. Using a direct comparison between antineutrino interaction rates of all cells, independent of any flux prediction, we find the data compatible with the null oscillation hypothesis. The best-fit point of the Reactor Antineutrino

Note: 29 pages, 33 figures, for supplemental data see http://doi.org/10.17182/hepdata.92323

Note: 29 pages, 33 figures, for supplemental data see http://doi.org/10.17182/hepdata.92323
Keyword(s): INSPIRE: * Automatic Keywords * | antineutrino: nuclear reactor | antineutrino: interaction | neutrino: sterile | antineutrino: energy spectrum | model: optical | n: capture | oscillation | anomaly | statistical | gadolinium | background | flux

inspirehep.net/collection/Conferences

HEPData

hepdata.net/record/ins1770821

HEPData | Search HEPData | Search

Browse all

Hide Publication Information

Improved Sterile Neutrino Constraints from the STEREO Experiment with 179 Days of Reactor-On Data

The STEREO collaboration

Almazán Molina, Helena , Bernard, Laura , Blanchet, Adrien , Bonhomme, Aurélie , Buck, Christian , del Amo Sánchez, Pablo , El Atmani, Ilham , Haser, Julia , Kandzia, Felix , Kox, Serge

No Journal Information, 2019

https://doi.org/10.17182/hepdata.92323

INSPIRE

Abstract (data abstract)
The STEREO experiment is a very short baseline reactor antineutrino experiment. It is designed to test the hypothesis of light sterile neutrinos being the cause of a deficit of the observed antineutrino interaction rate at short baselines with respect to the predicted rate, known as the Reactor Antineutrino Anomaly. The STEREO experiment measures the antineutrino energy spectrum in six identical detector cells covering baselines between 9 and 11 m from the compact core of the ILL research reactor. In this article, results from 179 days of reactor turned on and

Download All

Filter 4 data tables

Relative Spectra of IBD Events for Phase-I+II
10.17182/hepdata.92323.v1/t1
Data from Figure 30 – Relative comparison between the estimated rates of IBD events $A_{l,i}$ (for cell l and energy...

Exclusion and Sensitivity Contours
10.17182/hepdata.92323.v1/t2
Data from Figure 32 – STEREO exclusion and exclusion sensitivity contours at 90% C.L. for 179 days reactor-on (phase-I+II). A...

Delta Chi-squared Map
10.17182/hepdata.92323.v1/t3
$\Delta\chi^2$ map of phase-I+II data calculated by a raster-scan method. To be used in combination with the $\Delta\chi^2_{\text{crit},x}$...

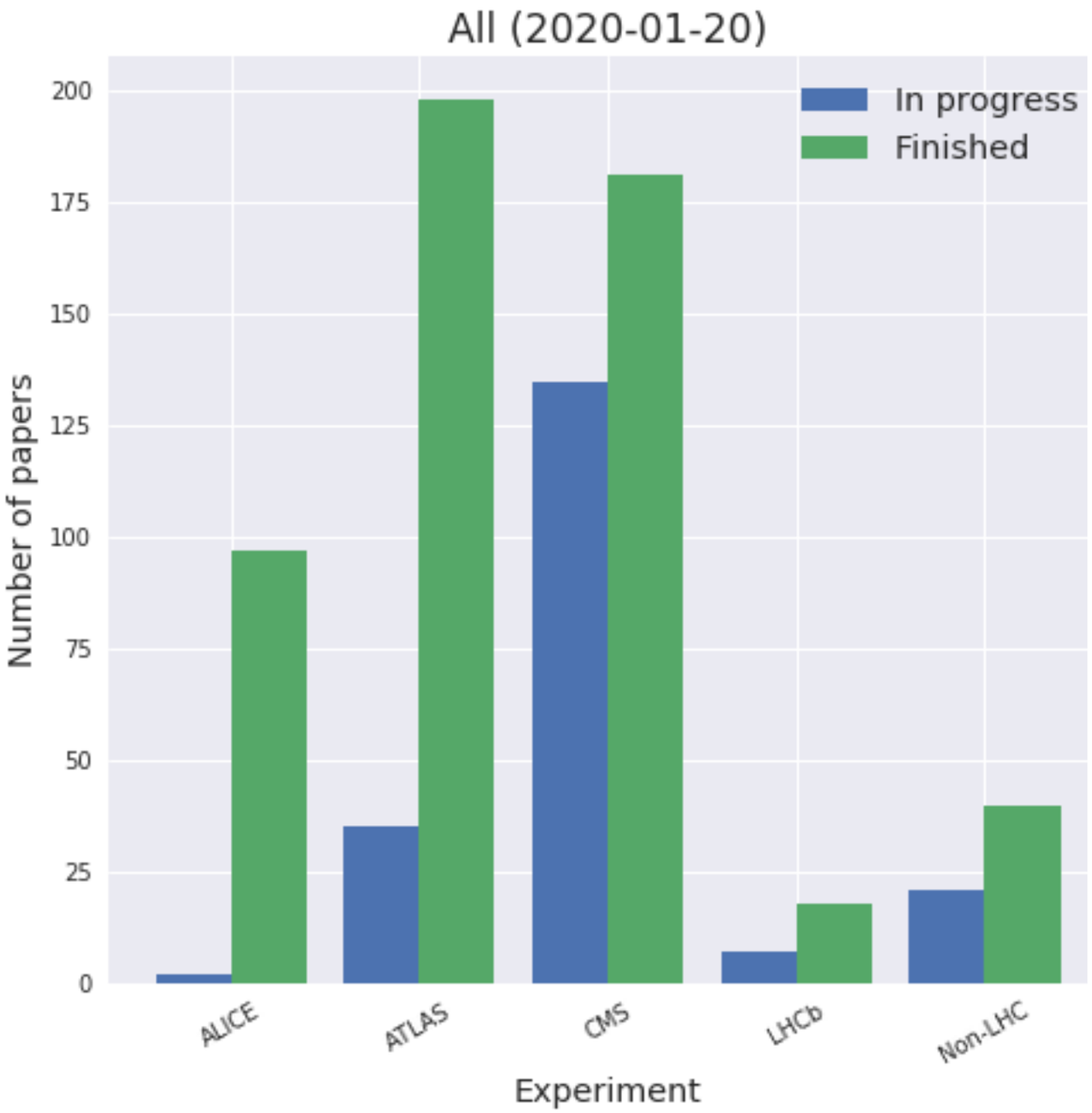
DATA AVAILABILITY

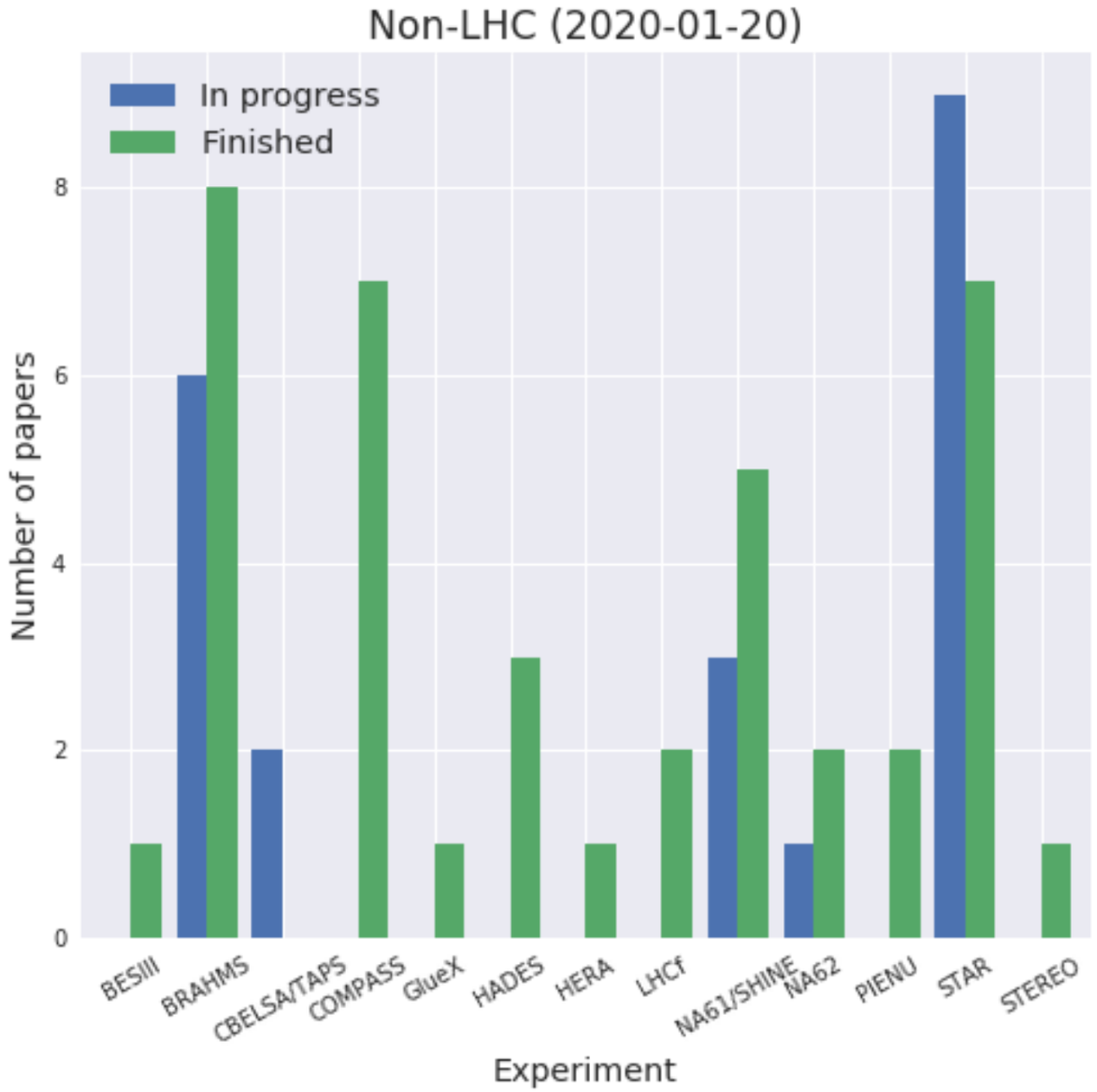
To allow inclusion of this work into global oscillation analyses and further works, we provide various results of our analysis in digitised form. These supplements can be found in reference [59].

[59] H. Almazán *et al.* (STEREO Collaboration), HEPData **92323** (2019).

- HEPData record released a few hours after arXiv paper made public.

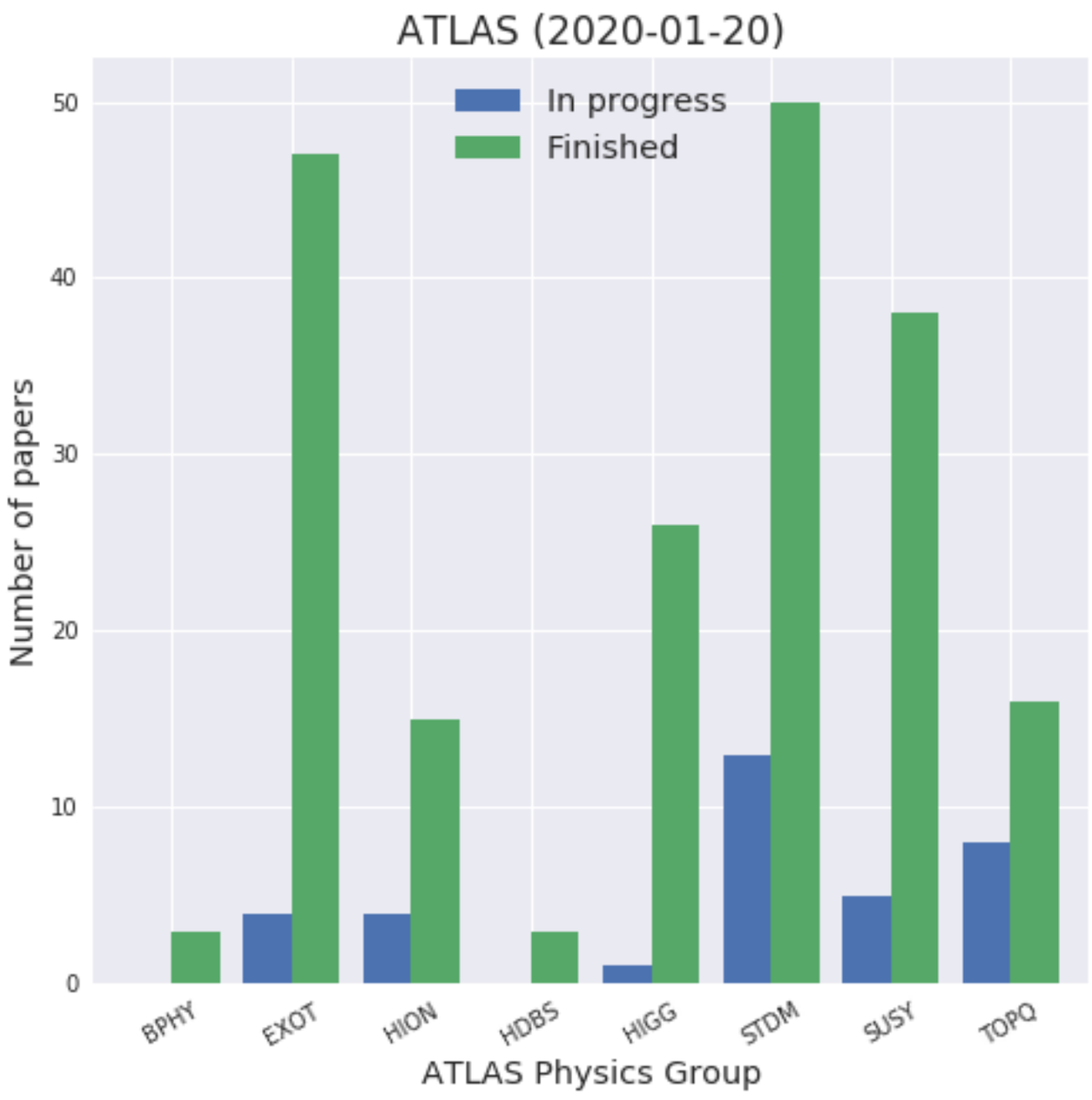
Submission system usage (01/2017-)

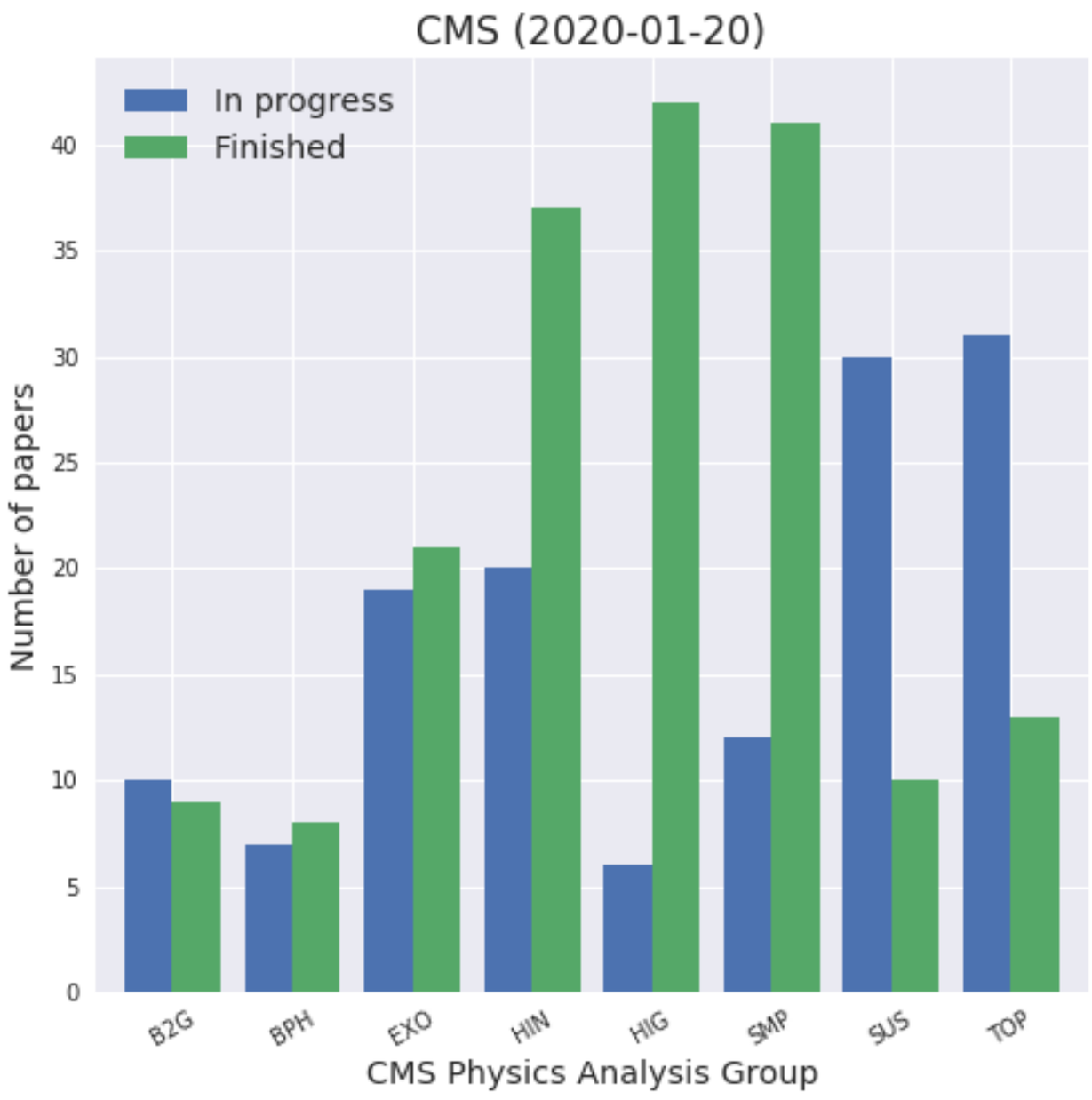




- Total of **534** finished submissions, comprising ALICE (97), ATLAS (198), CMS (181), LHCb (18), Non-LHC (40).

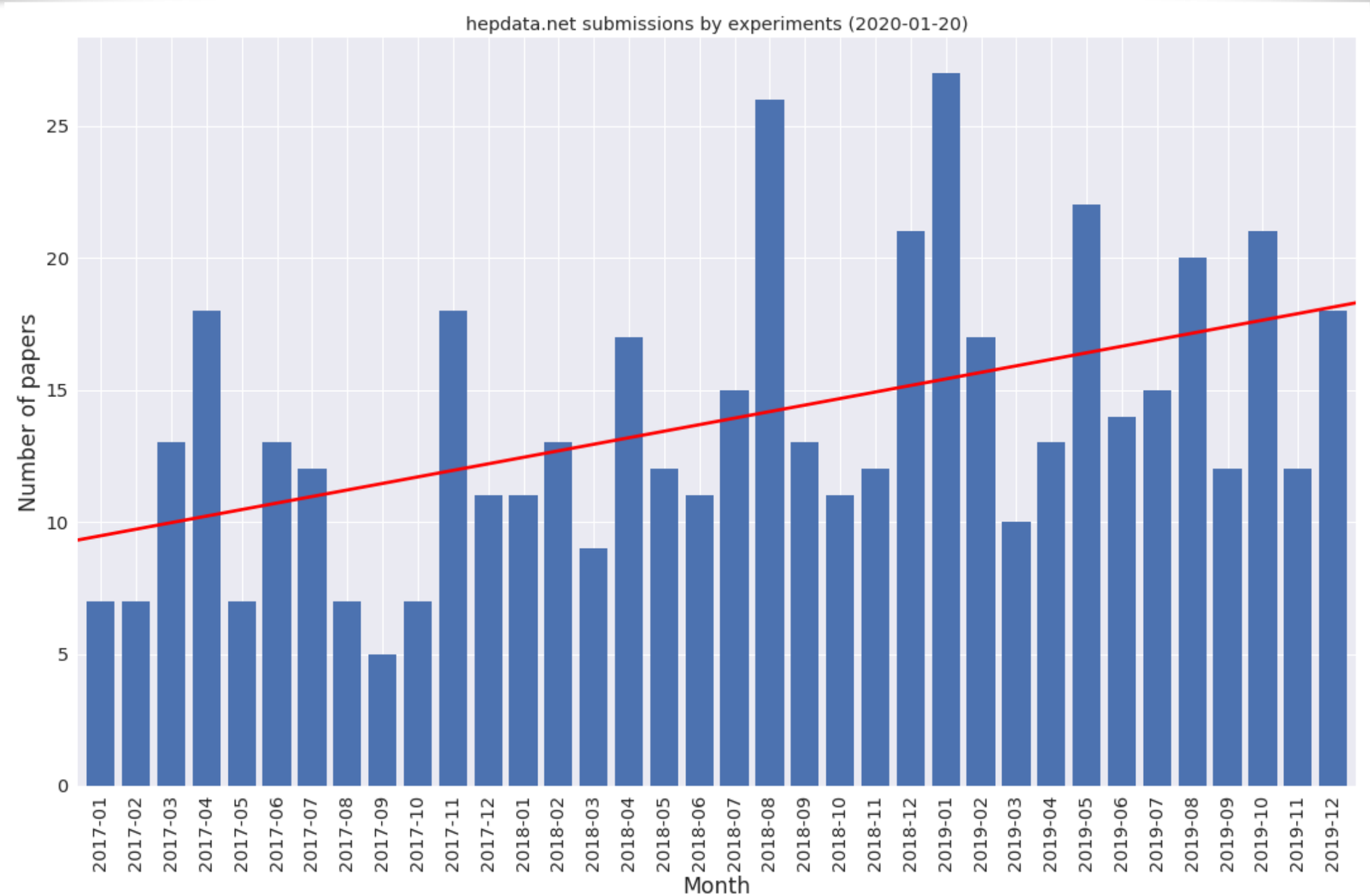
Submissions by ATLAS/CMS physics groups





- All **8** ATLAS and **8** CMS physics groups submit to HEPData.

Submissions per month from 2017 to 2019

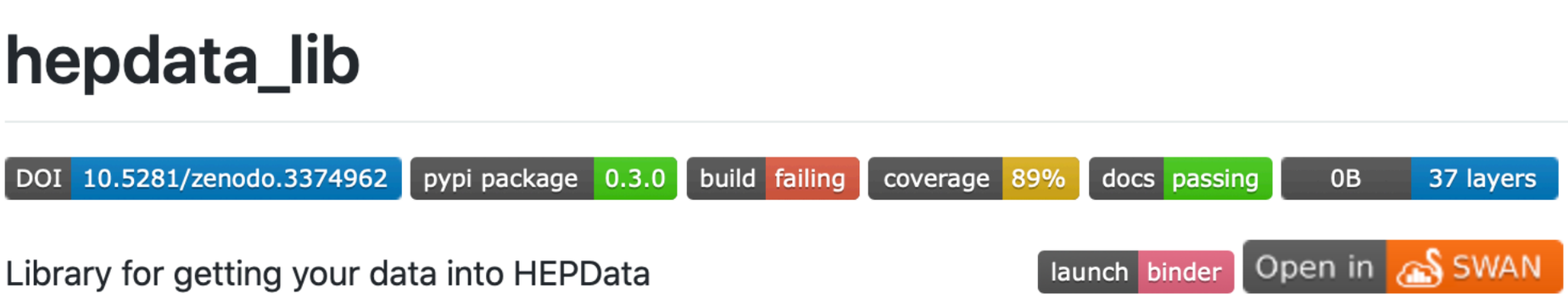


- Year-on-year growth: **125** in 2017, **171** in 2018, **201** in 2019.

Submission documentation

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

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



- Similar Python package by *Christian Holm Christensen*. https://gitlab.com/cholmcc/hepdata

HEPData / Rivet compatibility

- HEPData exports data in YODA format for Rivet.
- Divergence between Rivet data and HEPData export.
- Python scripts written in 2018 to assess compatibility: download `.yoda` from hepdata.net and call yodadiff.
- Rivet 2.6.1: 73 of 378 (**19%**) were compatible.
- Scripts rivet-diffhepdata and rivet-diffhepdata-all now included in official Rivet distribution.
- Rivet 3.1.0: 430 of 821 (**52%**) were compatible.
- Output from rivet-diffhepdata-all in web directory.
- Talk and discussions at Rivet workshop in May 2019.

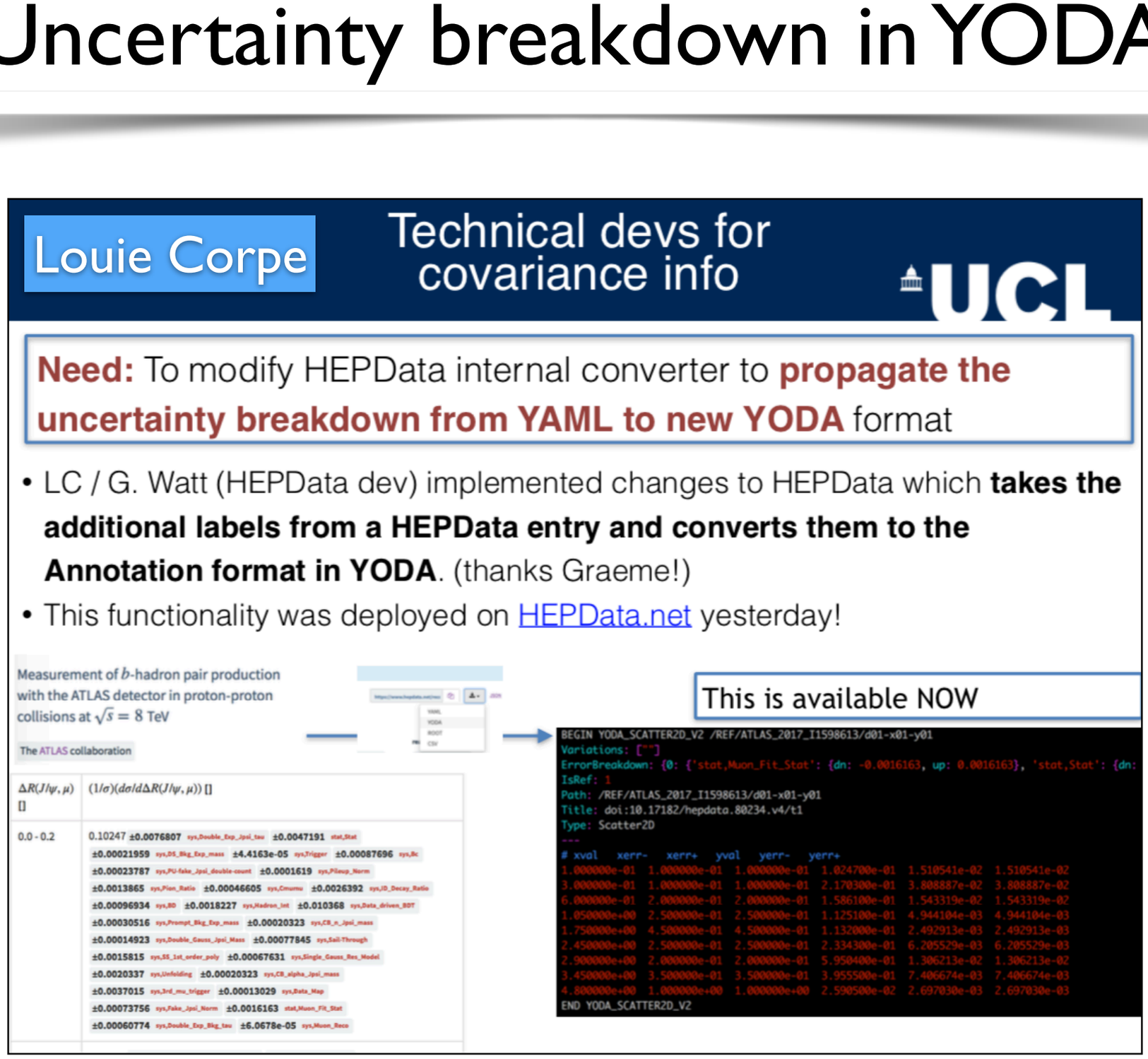
Thanks to Christian Gütschow.

Uncertainty breakdown in YODA

Louie Corpe — Technical devs for covariance info — UCL

Need: To modify HEPData internal converter to **propagate the uncertainty breakdown from YAML to new YODA** format

- LC / G. Watt (HEPData dev) implemented changes to HEPData which **takes the additional labels from a HEPData entry and converts them to the Annotation format in YODA**. (thanks Graeme!)
- This functionality was deployed on HEPData.net yesterday!



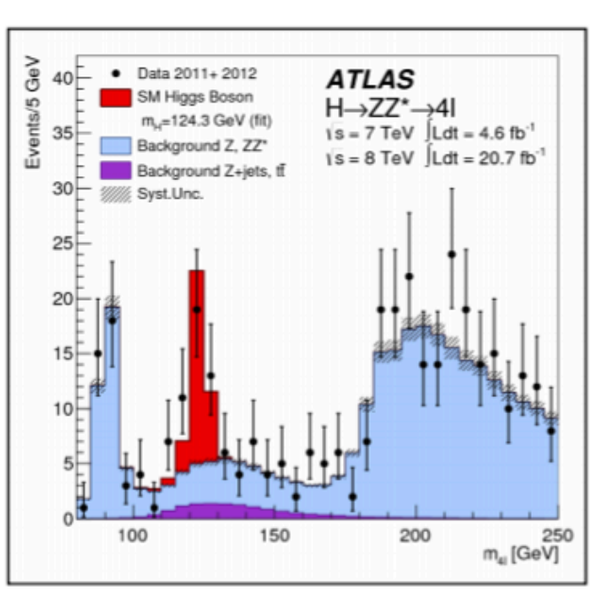
- Work with Louie Corpe (talk). New converter release deployed on 26th November 2018.

HistFactory

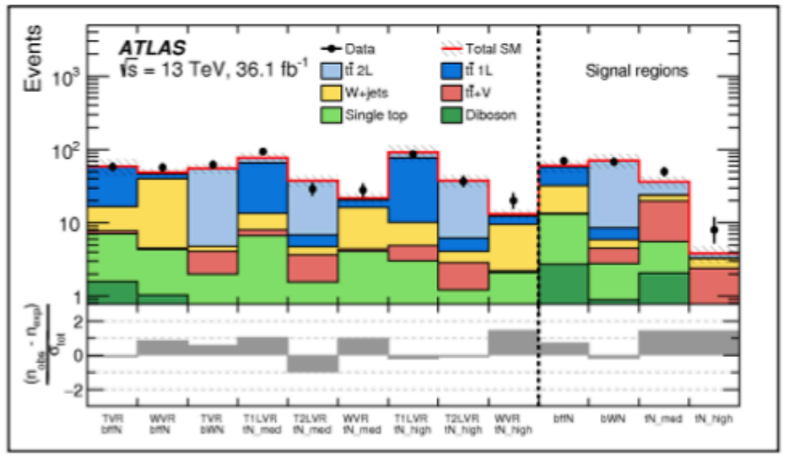
Enter HistFactory

Matthew Feickert (CHEP 2019)

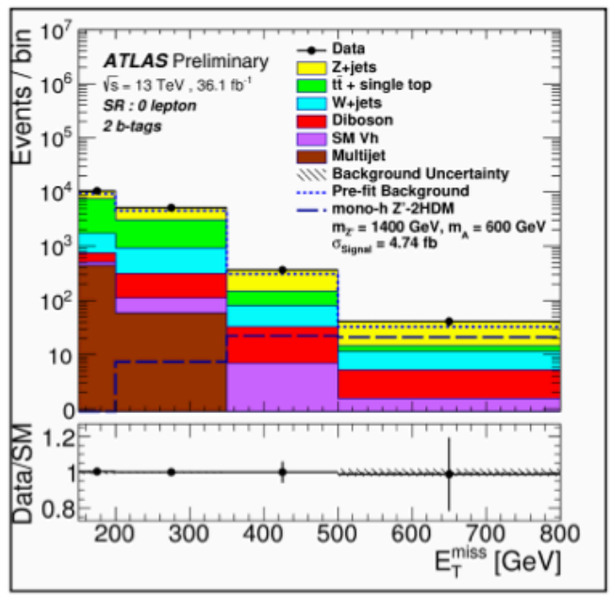
- A flexible p.d.f. template to build statistical models from binned distributions and data
- Developed by Cranmer, Lewis, Moneta, Shibata, and Verkerke (CERN-OPEN-2012-016)
- Widely used by the HEP community for standard model measurements and BSM searches



SM



SUSY



Exotics

- Based on simple ROOT histograms organised in an XML file.

pyhf likelihoods

Kyle Cranmer
Lukas Heinrich
Matthew Feickert
Giordon Stark

- Recent Python implementation (pyhf) of HistFactory.
- ROOT/XML workspace replaced by plain-text pyhf JSON.
- Two HEPData records released by ATLAS with pyhf JSON:
  https://www.hepdata.net/record/ins1748602 (2019-10-21)
  https://www.hepdata.net/record/ins1765529 (2019-12-06)
- Need to mint DOIs for local resource files and identify pyhf.
- **Native** support of pyhf JSON (see HEPData/hepdata#164):
  - ✦ Replace usual HEPData YAML data files with pyhf JSON.
  - ✦ Validate against JSON schema distributed with pyhf.
  - ✦ Develop appropriate *visualisation* and *conversion* tools.

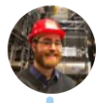
Publicity for likelihoods

New open release allows theorists to explore LHC data in a new way

The ATLAS collaboration releases full analysis likelihoods, a first for an LHC experiment

9 JANUARY, 2020 | By Katarina Anthony

Explore ATLAS open likelihoods on the HEPData platform (Image: CERN)



Matthew Feickert @HEPfeickert

Follow

The latest issue of @ATLASexperiment News highlights the first ever open publication of full likelihoods in high energy physics on @HEPData as well as the work that @lukasheinrich_, @kratsg, and I did to create the JSON schema for likelihoods that enabled this publication!

> **ATLAS Experiment** @ATLASexperiment
> What if you could test a new theory against LHC data? Better yet, what if the expert knowledge needed to do this was captured in a convenient format?…

8:22 am - 12 Dec 2019

- Articles linked from CERN home page and ATLAS News.

Funding from STFC

- Funded by UK **S**cience & **T**echnology **F**acilities **C**ouncil (**STFC**), with previous 3-year PPGP(E) grant until 09/2019.
- Last 3-year grant application submitted in February 2018 with outcome in April 2019. Application unsuccessful, but stop-gap of 6 months (to be extended until 09/2020).
- Awarded 2 FTE for *Project Manager* + *Software Engineer*.
- **STOP PRESS**: Yesterday STFC informed us that they will add HEPData funding to the upcoming PPGP(Theory) Consolidated Grant (10/2020-09/2023) for the IPPP.

Current STFC grant (10/2019-03/2020)

- Promised 04/2020-09/2023, but funding not yet received.
- Ended *Data Entry Assistant* (Joanne Bentham, 1995-2018).
- Awarded 2 FTE for *Project Manager* + *Software Engineer*.

- How to recruit a *Software Engineer* to fill a short contract?
- Recent growth of Research Software Engineering (RSE): UK RSE Association and Society of RSE (launched 2019).
- New group of 3 RSEs within ARC at Durham University.
- Alison Clarke working 0.4 FTE on HEPData from 11/2019.

Alison Clarke's contributions

Improvements to code for main web app:

1. Display message to user when table fails to load (PR).
2. Use a different Elasticsearch index for tests (PR).
3. Increase Python test coverage from 68% to 76% (PR).
4. Resurrect Selenium end-to-end tests via SauceLabs (PR).
5. Ensure users verify their email addresses to log in (PR).
6. Porting the codebase from Python 2 to Python 3 (Issue).

- **Excellent progress** in only 2 months at 0.4 FTE.

Research Software Healthcheck

- Service provided by the Software Sustainability Institute:

The Research Software Healthcheck is a new, free evaluation service from the Institute to help your research software reach its potential and make it more sustainable.

A lightweight evaluation of your software will provide a set of concrete recommendations for you to help prepare it for future challenges and opportunities. The evaluation can be tailored to focus on any aspects of particular concern to you.

Applications from any UK institution and research area are eligible.

- Generated report from online sustainability evaluation.
- Healthcheck application submitted on 12th November 2019.
- Skype with Steve Crouch and James Graham on 10/01/2020.
- Outcome of HEPData software evaluation expected soon.

Past hepdata.net operational issues

- Resolve EOS upload issues by unpacking archive in a temporary directory, then move to EOS using '`xrdcp`'.
- Downtime still usually caused by EOS disk inaccessibility.
- January 2018: OpenStack cluster for converter died.
- March 2018: latest Invenio v3 packages broke build. Pinned to older versions via requirements.txt (Issue).
- Needed to increase EOS quota from 1 M files to 2 M files in April 2018 then to 4 M files in July 2019 (Issue to clean up).
- September 2018: exceeded 1000 emails/month quota from SMTP2GO. Use CERN mail gateway instead (Issue closed).

Future hepdata.net operational tasks

- Management of Puppet configuration for Virtual Machines.
- VMs now 4 years old and will be retired by March 2020.
- Create replacement VMs with a Python 3 environment.
- Set up a **test** server in addition to the production server.
- Configure a new Sentry instance for error monitoring.
- Set up a framework to monitor site usage from Nginx logs.
- Renew SSL certificates and firewall openings when required.
- Manage the cluster running the converter Docker container.
- Monitor the operation of the VMs using Nagios.

HEPData and INSPIRE

- Discussed IPPP Durham accession to INSPIRE Collaboration since 2015, but formal collaboration agreement never signed.
- Revisit now that IPPP Durham has stable funding for HEPData.
- CERN contribution in draft collaboration agreement:
  - Maintenance of CERN hosting of the HEPData service
  - Support troubleshooting and debugging HEPData service operations
- HEPData operations much in common with INSPIRE service.
- Original motivation of moving HEPData to Invenio v3 was to allow closer integration with INSPIRE beta: not yet achieved.
  - Need to link HEPData records to new INSPIRE beta records.
  - Track DOI citations to HEPData records with INSPIRE beta.

Conclusions

Email: info@hepdata.net

Follow @HEPData

- Achievement to keep hepdata.net alive and functional for last three years without support of a dedicated Software Engineer.
- **Steady growth** in number of users and data submissions.
- Backlog of issues: https://github.com/HEPData/hepdata/issues
- Alison Clarke (RSE) now working 0.4 FTE since 11/2019.
- **STOP PRESS**: STFC awarded HEPData funding to 09/2023.
- Future staffing recommendation assuming funding for 2 FTE:
  - Project Manager @ 1 FTE (Durham, IPPP)
  - Development Software Engineer @ ~0.5 FTE (Durham, ARC)
  - Operations Software Engineer @ ~0.5 FTE (CERN)