HEPData status

Graeme Watt (Database Manager)

HEPData advisory board meeting

IPPP Durham, 24th November 2017

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 more than 8500 experimental **HEP** (“`hep-ex`”) papers.
- “**Level 1**” according to DPHEP classification [1205.4667], e.g. numbers for final data points plotted in publications.
- Based in **I**nstitute for **P**article **P**hysics **P**henomenology (**IPPP**) at Durham University (UK), going back to 1970s.
- Recent transition to hepdata.net site, hosted at CERN.
- Funded by UK **S**cience & **T**echnology **F**acilities **C**ouncil (**STFC**). Current grant to support two staff until 2019.
- Grant application for 2019 to 2022 due in February 2018.

HEPData advisory board

Jeppe Andersen (IPPP Durham)
Andy Buckley (Glasgow, Rivet)
Kyle Cranmer (NYU, generic users)
Sünje Dallmeier-Tiessen (CERN, INSPIRE)
Ulrik Egede (Imperial, LHCb)
Keith Ellis (IPPP Durham)
Henning Flaecher (Bristol, CMS)
Daniel Maître (IPPP Durham)
Salvatore Mele (CERN, INSPIRE)
Bill Murray (RAL/Warwick, ATLAS)
Enrico Scomparin (Torino, ALICE)
Matthew Wing (UCL, non-LHC experiments)

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

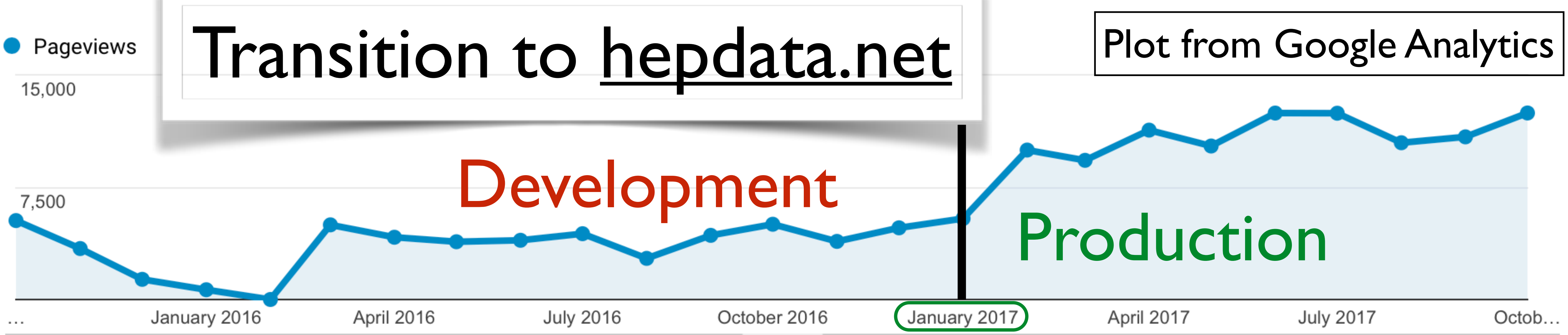
- New members:
  - Andy Buckley
  - Daniel Maître

HEPData staff

- **Database Manager:** Graeme Watt (2013-)
- **Data Encoding Assistant:** Joanne Bentham (1995-)

- Other personnel changes since last meeting:
  - **Previous Database Manager:**
    Mike Whalley (Durham, 1982-2016) ➔ Graeme Watt
  - **Lead Developer for hepdata.net:**
    Eamonn Maguire (CERN, 2015-2016) ➔ Graeme Watt
  - **Principal Investigator:**
    Frank Krauss (Durham, 2014-2017) ➔ Keith Ellis

Transition to hepdata.net



- Complete rewrite of HEPData software as an overlay on Invenio digital library framework, but heavily customised.
- Decision after last meeting to move from Invenio v2 to v3.
- Improvements and fixes to code base throughout 2016/17.
- Bulk migration of all records from old site in October 2016.
- Internal use of new submission system towards end of 2016.
- External data submissions from January 2017 onwards.
- Old site (hepdata.cedar.ac.uk) now officially superseded.

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

arXiv:1704.05473v1 [hep-ex] 18 Apr 2017

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

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

[arXiv:1704.05473]

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

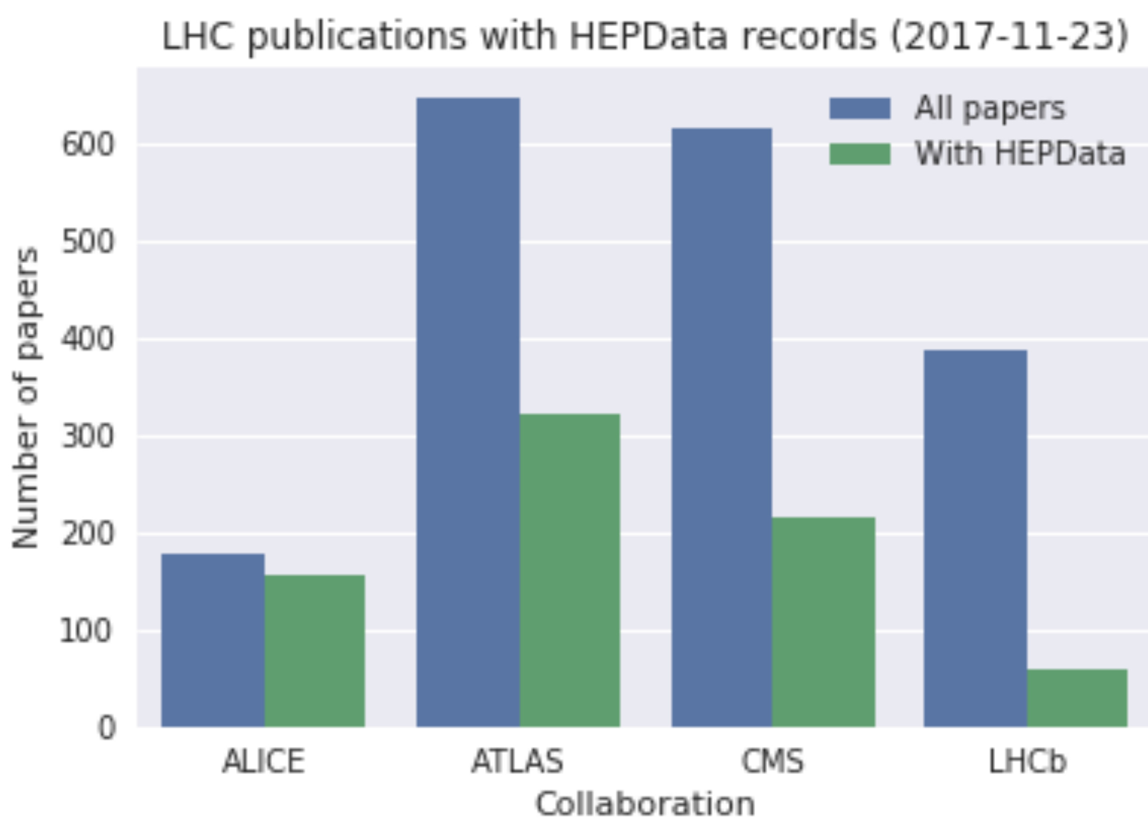
HEPData is funded by a grant from the UK Science and Technology Facilities Council. The DOI minting originates from the THOR project, funded by the European Commission under the Horizon 2020 programme. We are indebted to Mike Whalley for his dedicated 34 years of service as Database Manager for previous 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

Links from inspirehep.net

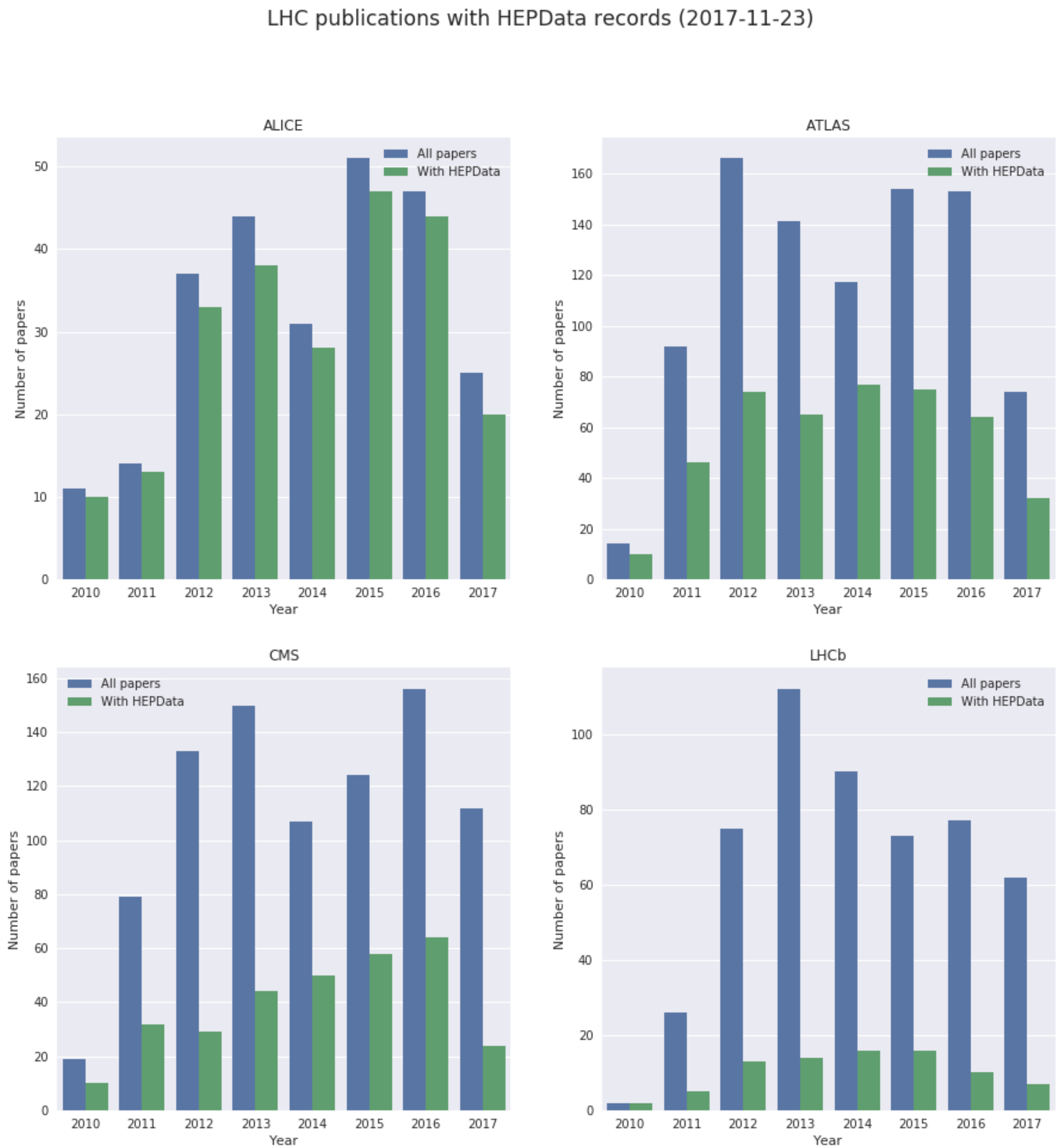
- INSPIRE is currently also being rewritten using Invenio v3.
- Legacy INSPIRE records updated to link to hepdata.net.
- Search INSPIRE for HEPData records with "035:hepdata".
- LHC publications with HEPData (Jupyter Notebook):



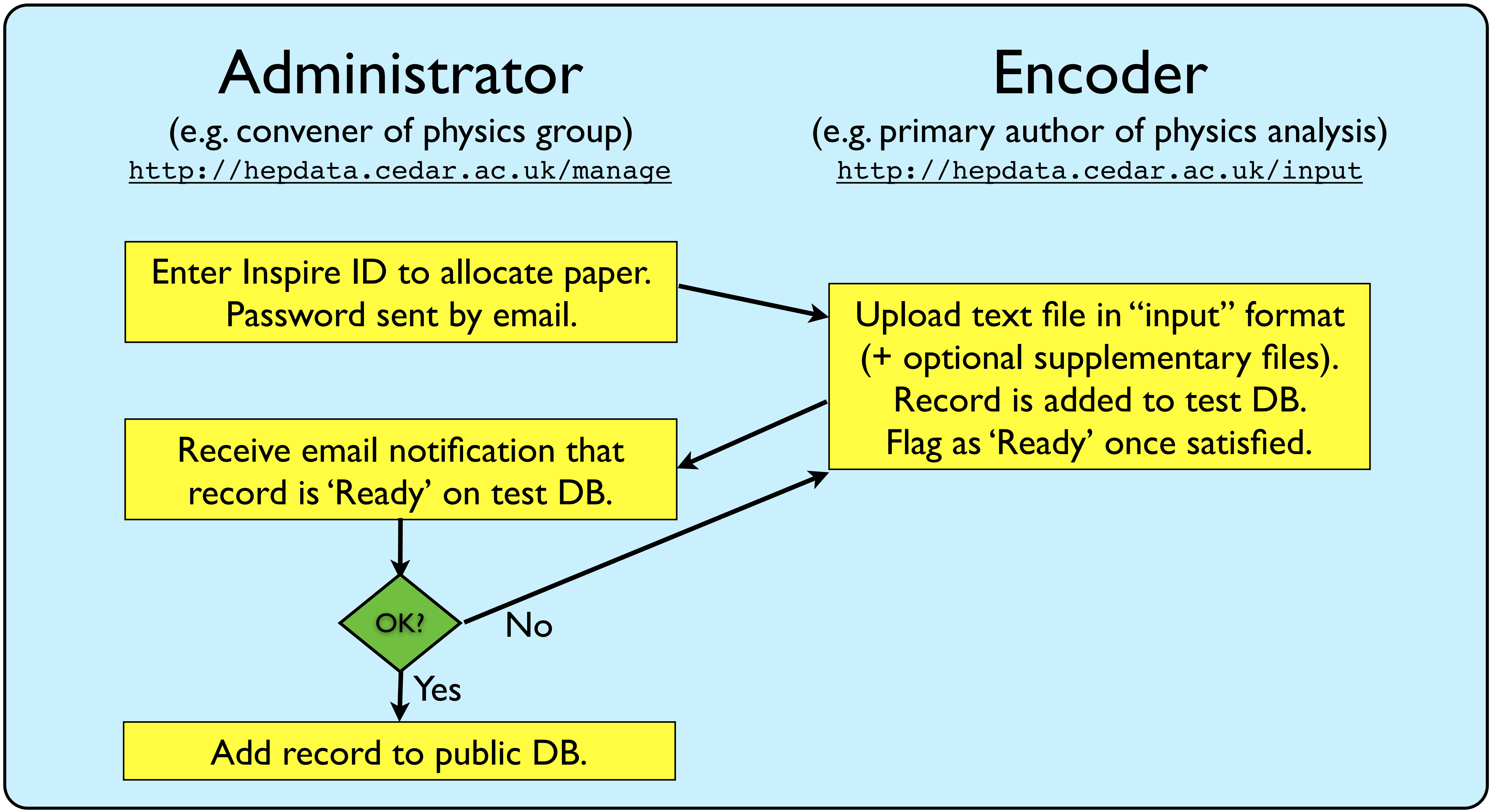
INSPIRE search query:
- `hep-ex` or `nucl-ex`
- Published in a journal
- Not conference paper

ALICE: 88%
ATLAS: 50%
CMS: 35%
LHCb: 15%

Coverage of LHC publications by year

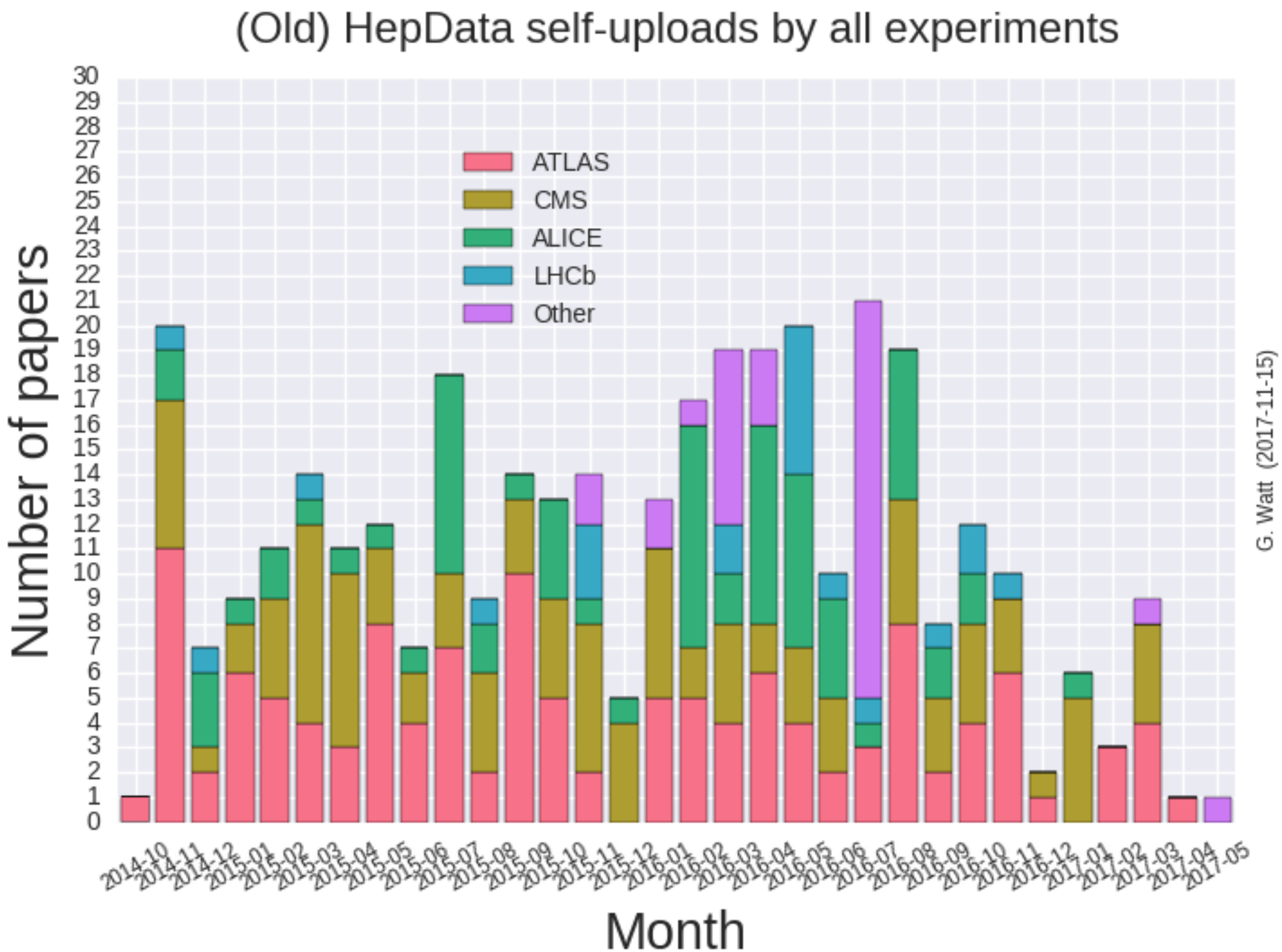


Old submission system workflow (10/2014-05/2017)



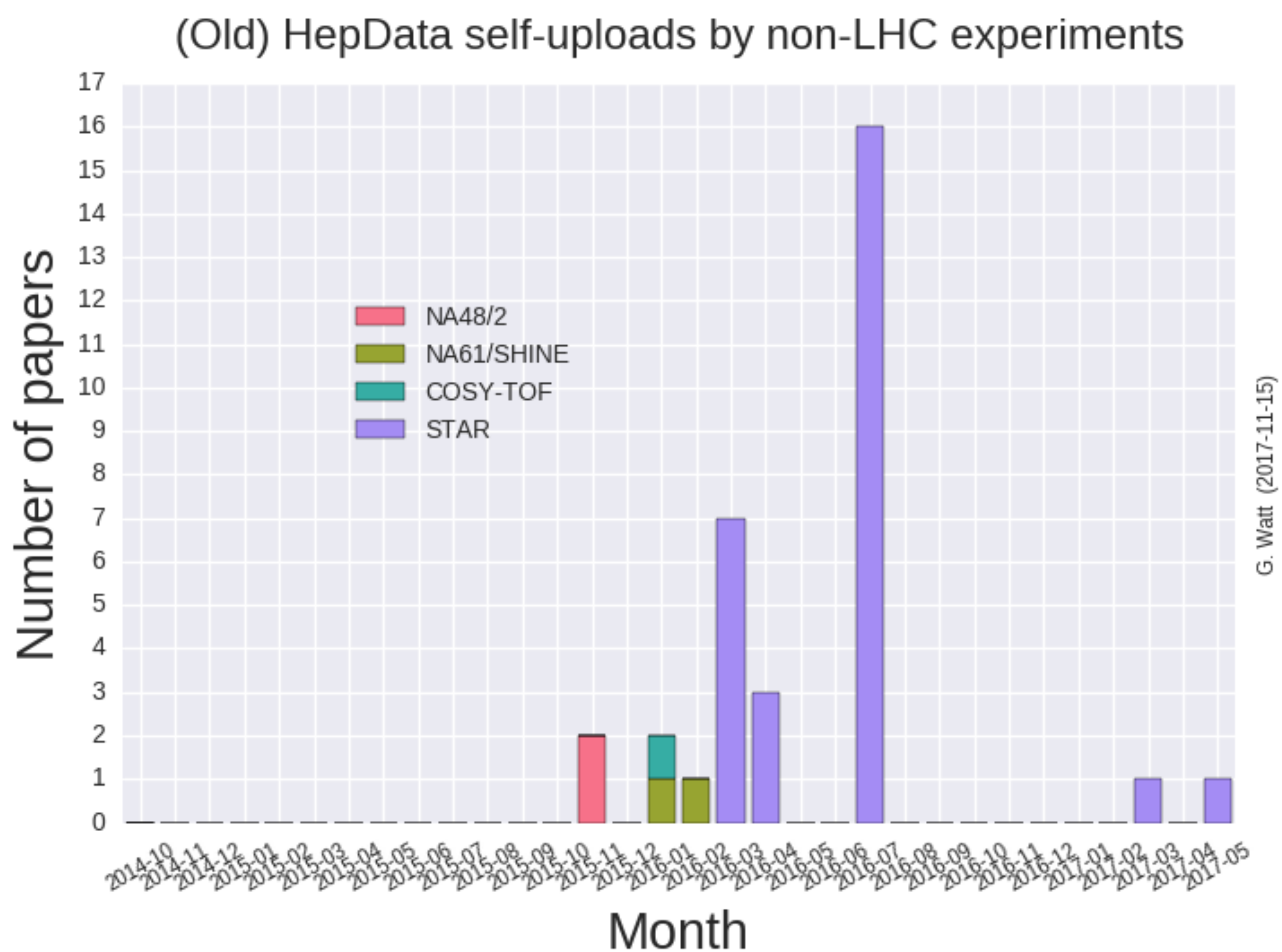
- First paper added by ATLAS SUSY group on 28th October 2014.

Old submission system usage (10/2014-05/2017)



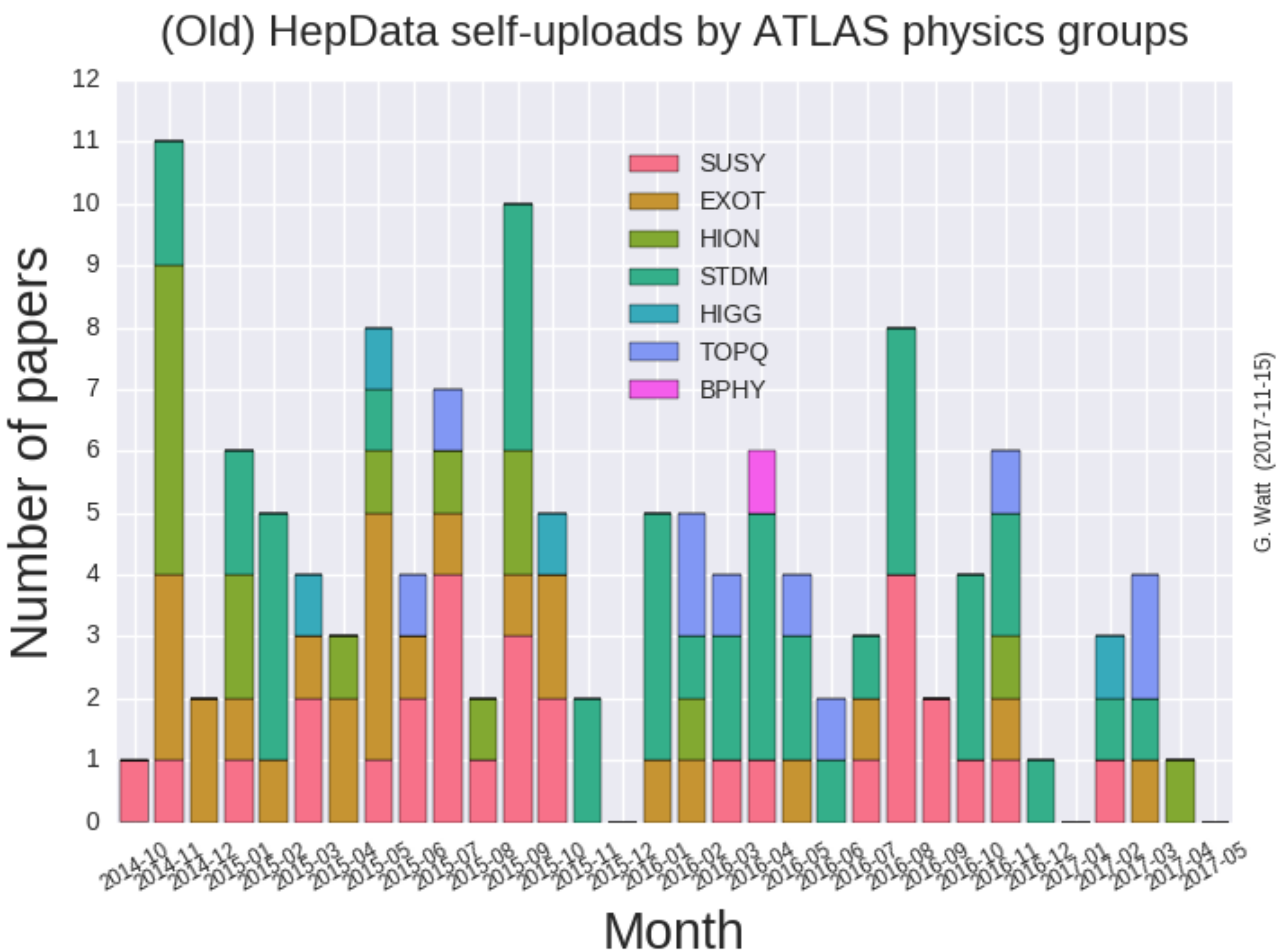
- Total of 355 papers: ATLAS (128), CMS (102), ALICE (71), LHCb (21), Other (33).

Old submission system usage by non-LHC



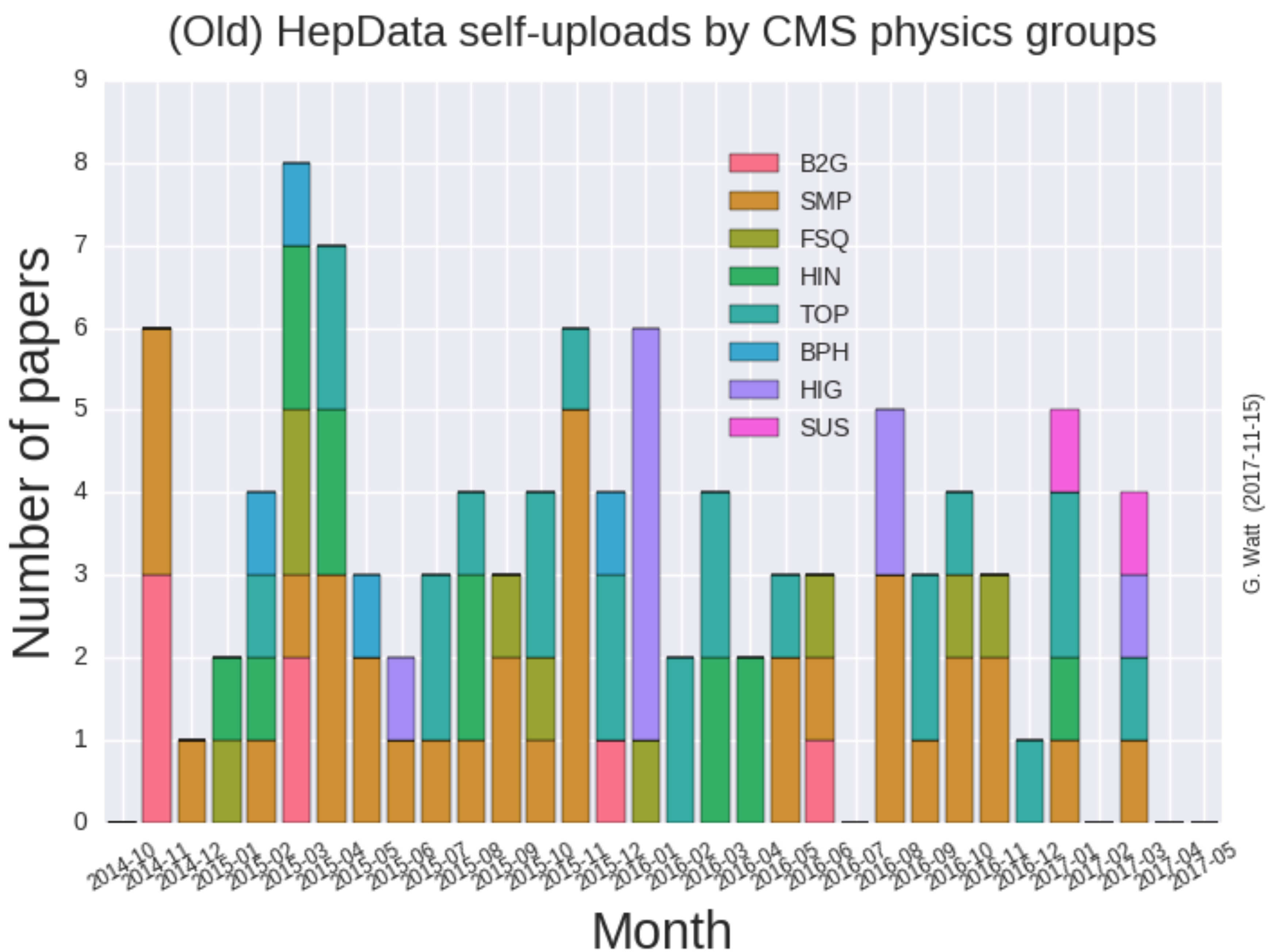
- Total of 33 non-LHC papers: NA48/2 (2), NA61/SHINE (2), COSY-TOF (1), STAR (28).

Old submission system usage by ATLAS



- Total of 128 ATLAS papers: SUSY (30), EXOT (25), HION (16), STDM (42), HIGG (4), TOPQ (10), BPHY (1).

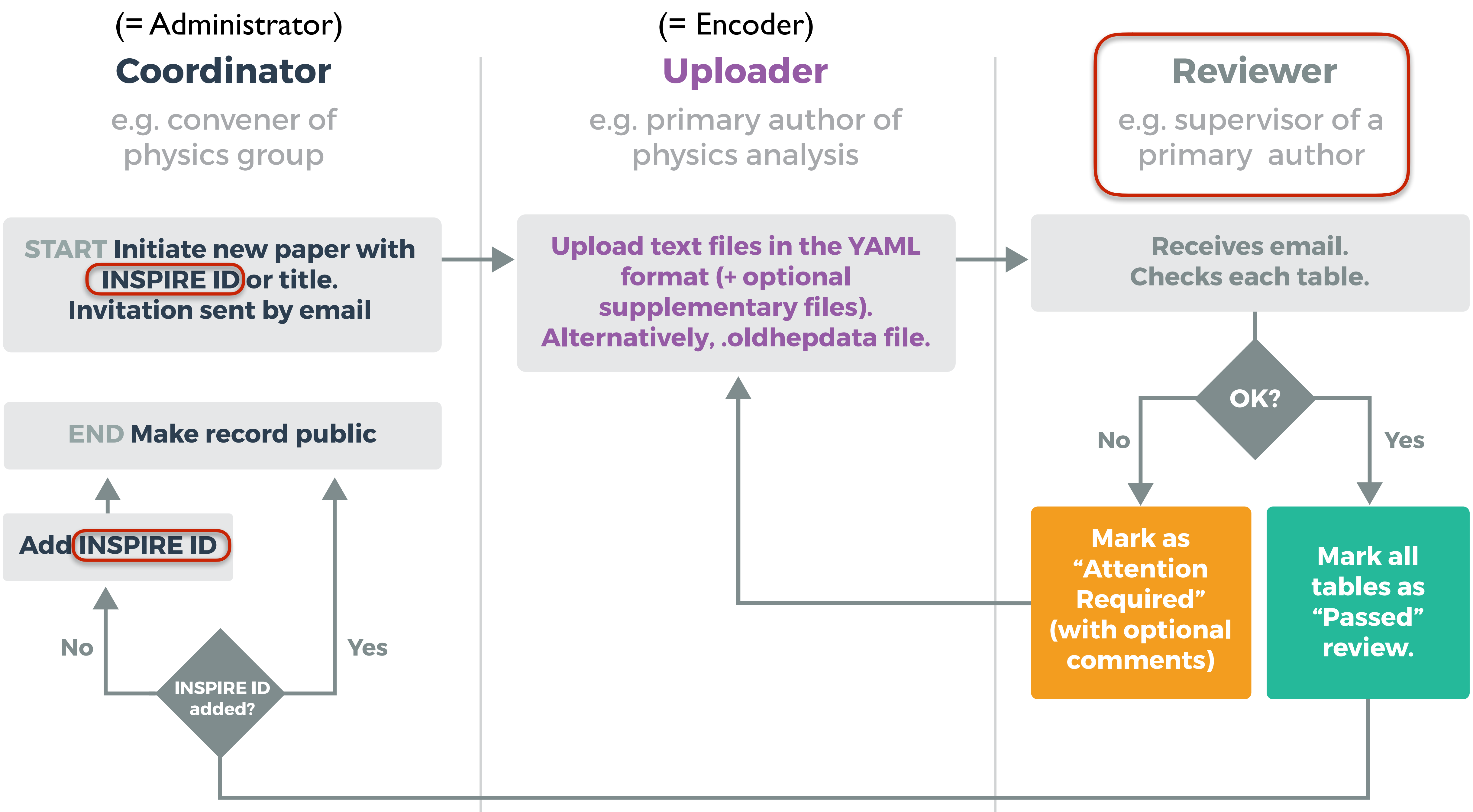
Old submission system usage by CMS



- Total of 102 CMS papers: B2G (7), SMP (35), FSQ (9), HIN (13), TOP (23), BPH (4), HIG (9), SUS (2), EXO (0).

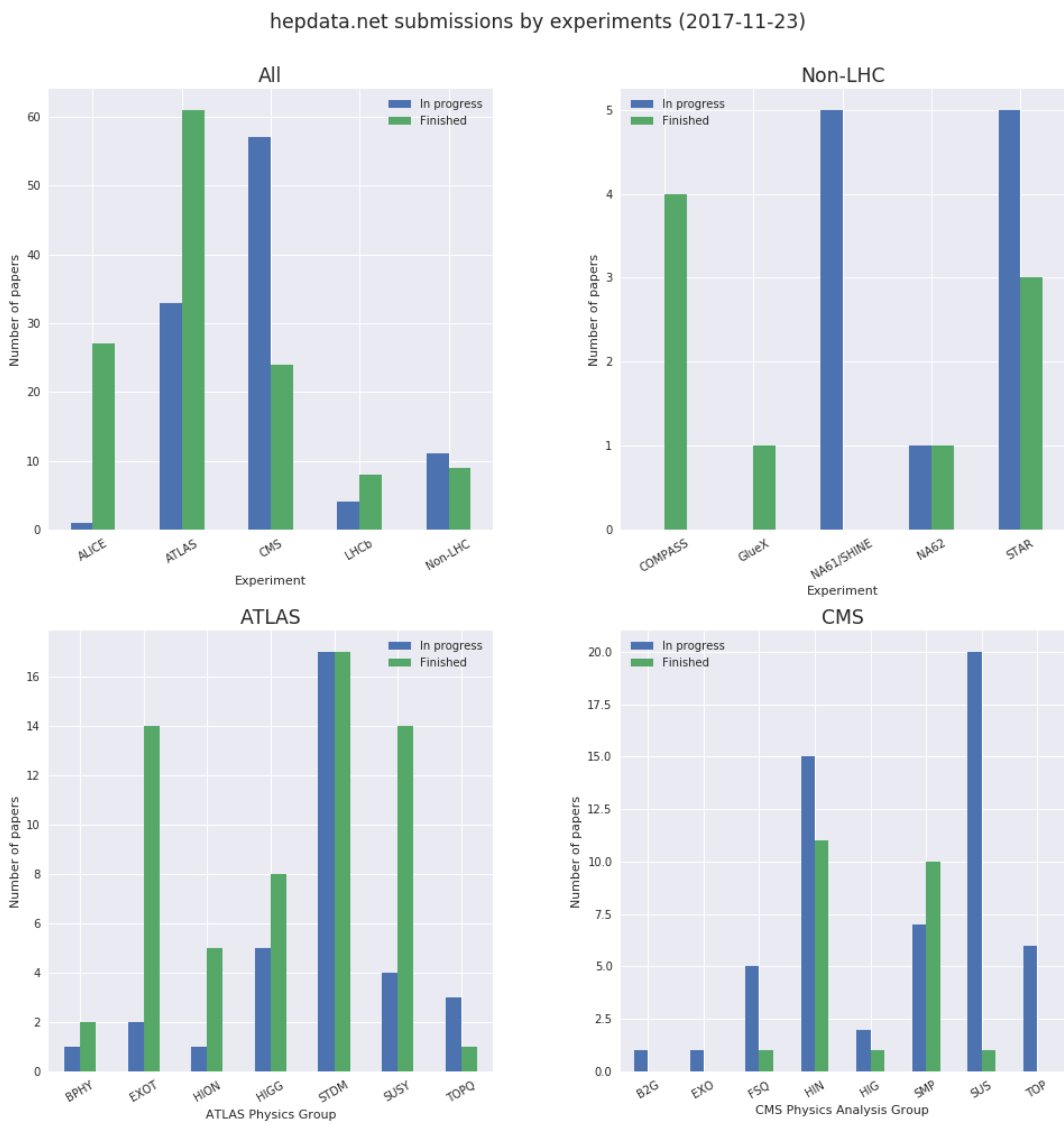
New submission system on hepdata.net

https://hepdata.net/submission



- Also a Sandbox for any user to test uploads without special privileges.

New submission system usage (01/2017-)



437 users created an account (as of 23/11/2017)

- Numerous improvements and fixes made to code and documentation based on large amount of feedback received.
- Total of **129** finished submissions:
  - ALICE (27)
  - ATLAS (61)
  - CMS (24)
  - LHCb (8)
  - Non-LHC (9)
- **24**/129 with version ≥ 2

Data encoding assistant

- Database Manager selects "`hep-ex`" arXiv papers older than 1 year, that contain tables of numbers in the LaTeX source. Decides which tables to extract, how to organise them, and specifies keywords etc.
- Assistant prepares input file from LaTeX using a text editor (`NEdit`).
- Database Manager checks input file and adds record to public database.
- **59** papers completed in 2016. **29** papers completed so far in 2017: CMD-3, KEDR, Belle, NA61/SHINE, ATLAS, Super-Kamiokande, BaBar, D0, MINERvA, BESIII, ATLAS, ATLAS, BESIII, MINERvA, CDF, KLOE-2, SND, MINOS, T2K, BESIII, CDF, CMS, CMS, NA61/SHINE, ANTARES, SLAC E140, CMD-3, MINERvA, CMS.
- Recent training in Python, but manipulating LaTeX source usually quicker.
- Need to evaluate whether this rôle can continue to be justified in future.

Data releases from neutrino experiments

[IPPP/NuSTEC topical meeting on neutrino-nucleus scattering]

- **MiniBooNE** (data releases for 13 papers): http://www-boone.fnal.gov/for_physicists/data_release/ No existing HEPData records.

- **T2K** (data releases for 20 papers): http://t2k-experiment.org/results HEPData: 1329784, 1394549, 1408741, 1421157, 1465650.

- **MINERvA** (data releases for 16 papers): http://minerva.fnal.gov/data-release-page/ HEPData: 1351216, 1393997, 1405301, 1416818, 1446753.

Numbers are INSPIRE IDs.

- NUISANCE framework contains a collection of data.

- Tried to engage to submit to HEPData, but without uptake.

HEPData beyond HEP?

- Visit from Maria Grazia Pia (INFN Genova) and students in December 2016 to discuss inclusion of data relevant to particle transport codes like Geant4.
- First external tests of hepdata.net submission system. Students wrote report.
- Presentation and conference paper at 2017 IEEE Nuclear Science Symposium.
- Fairly significant software development would be needed to extend database:
  - Introduce searchable keywords specific to nuclear/atomic physics.
  - Remove requirement of associated publication in INSPIRE.
- Interest in bulk migration of existing nuclear physics databases (e.g. ENDF).
- Would need to be careful that significant rewrite of code and expansion of database does not compromise performance of site for existing HEP users.
- Alternative: create a dedicated data portal (e.g. HEPData fork, Invenio, CKAN).

Open issues

- Many bug reports and feature requests with transition to new hepdata.net site. Many issues resolved, but also many remaining, e.g. https://github.com/HEPData/hepdata/issues
  - #59 editing: allow an admin user to edit records within a browser
  - #60 submission: build a tool to help a data uploader to select appropriate keywords
  - #100 records: allow users to update publication information
- Project suffers from lack of paid software engineering support: seek to improve for the next funding period.

Summary

Email: info@hepdata.net

Follow @HEPData

- Transition to rewritten **HEPData** site (hepdata.net) hosted at CERN. Many improvements and new features.
- New submission system successfully adopted by experiments.
- Minimal technical support from CERN since departure of Eamonn Maguire in 10/2016: need to clarify responsibilities.
- Project suffers from lack of paid software engineering support: seek to improve for the next funding period (2019-2022).
- Staffing: physicist?, data encoding assistant?, software engineer?
- Project priorities for next funding period? Input welcome!