Graeme Watt (IPPP Durham)

IPPP Computing meeting
Thursday 29th February 2024

https://hepdata.net

Email: info@hepdata.net
Forum: hepdata-forum.cern.ch

Follow @HEPData

Code: https://github.com/HEPData

What is HEPData?

- Unique *open-access* repository for tabular high-level **data** from more than 10k **HEP** publications (130k data tables).
- **FAIR** data: **F**indable, **A**ccessible, **I**nteroperable, **R**eusable.
- *Complementary* to other HEP information providers, e.g. INSPIRE-HEP (literature), PDG (particle properties), CERN Open Data (event-level data), Zenodo (files).
- Historically based at Durham University (UK) from 1970s.
- Transition in 2017 to hepdata.net site, hosted at CERN. Partnership with CERN Scientific Information Service. J. Phys.: Conf. Ser. 898 102006 [arXiv:1704.05473]
- Old IPPP-hosted HepData server was shut down in 2022.

HEPData Infrastructure

- All provided by CERN IT with support from CERN SIS.
- Migration in 2020 from Puppet VMs to Docker/Kubernetes.
- Kubernetes configuration specified in private GitHub repo.
- Argo CD for monitoring and Sentry for error tracking.
- Shared CephFS storage for 1.3M data files (110 GB).
- Database On Demand (DBOD): PostgreSQL v14.6 (2.1 GB).
- OpenSearch v2.11.1 cluster indexes metadata for searching.
- Separate **QA** environment for testing before *production*.
- Discourse instance for Forum: hepdata-forum.cern.ch

HEPData Staff in Durham

Funded by STFC and included in IPPP grant since 10/2020.

1. **G.W.** (10/2013 - present, former pheno researcher):
   Manage project, provide user support and troubleshoot problems, develop software and documentation, monitor operation and usage of service, supervise *Software Engineer*.

2. *Software Engineer* from Advanced Research Computing works on development tasks under supervision by G.W.:

   - **Alison Clarke** (Senior RSE, 11/2019 - 05/2022)
   - *Zeynep Akı* (Assistant RSE, 04/2022 - 05/2022)
   - **Jordan Byers** (Assistant RSE, 06/2022 - present) supported by *Samantha Finnigan* (Senior RSE, 05/2023 - present)

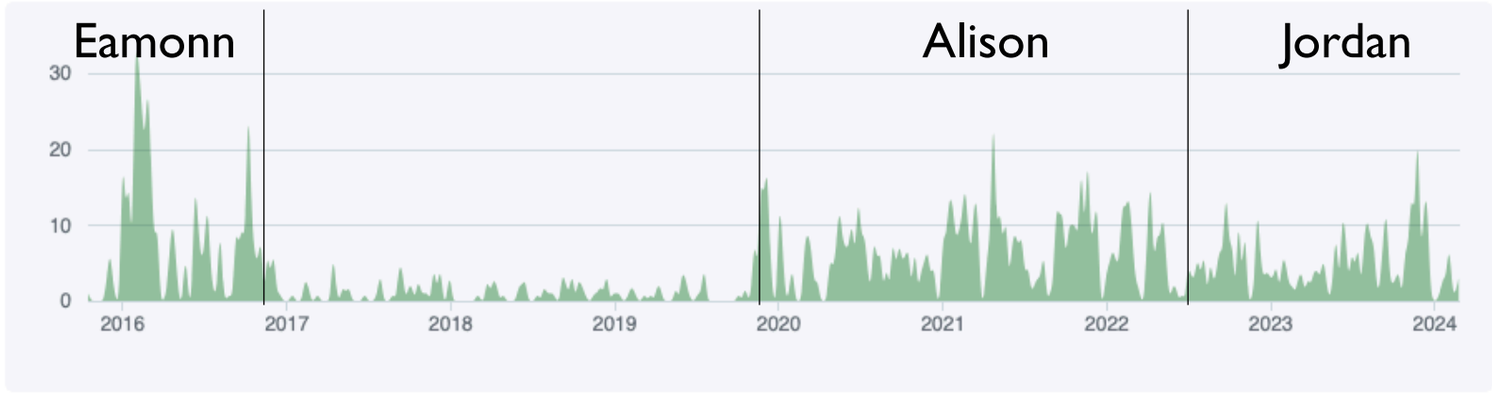
RSE Contributions

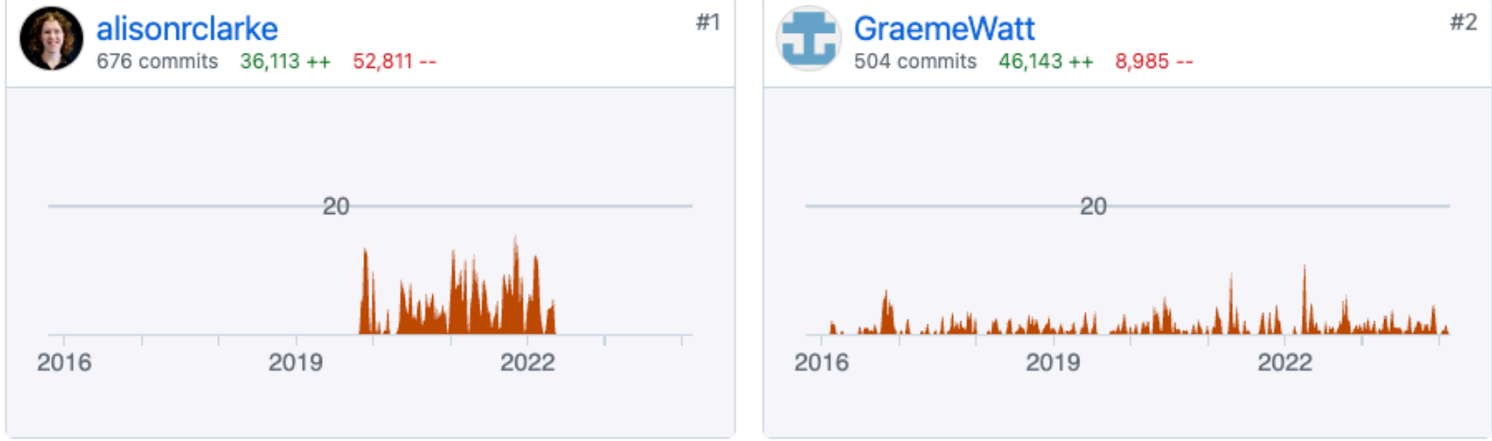
Oct 18, 2015 – Feb 28, 2024

Contributions: Commits

Contributions to main, excluding merge commits

https://github.com/HEPData/hepdata/graphs/contributors





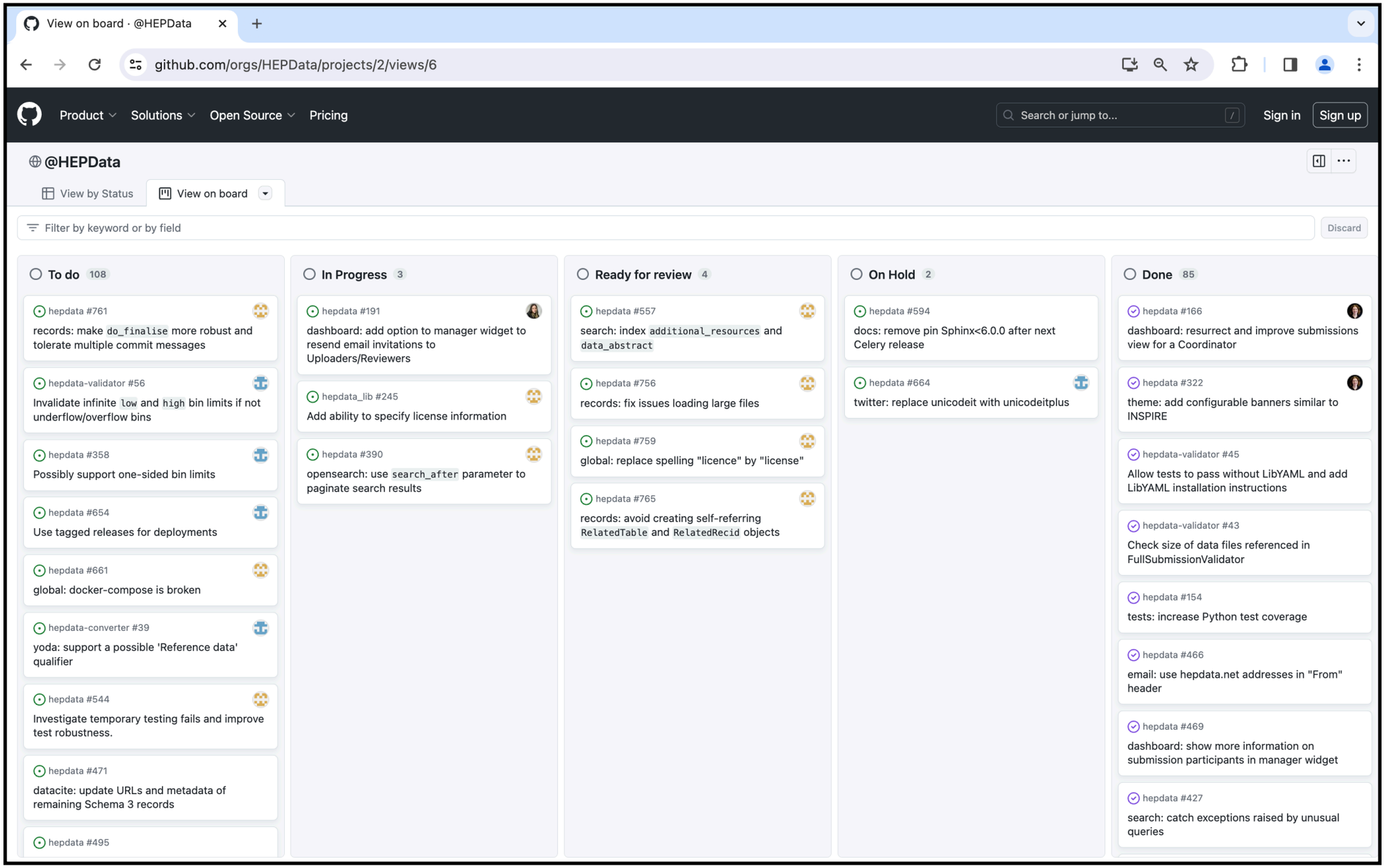
https://github.com/HEPData

- `hepdata`: main web application (Python, JavaScript, HTML)
- `hepdata-validator`: JSON schema and validation code
- `hepdata-submission`: documentation and examples
- `hepdata-converter`: YAML to CSV/ROOT/YODA
- `hepdata_lib`: helps transform text/ROOT files to YAML
- `hepdata-cli`: search/download/upload from CLI or API
- `miscellaneous`: Jupyter notebooks for various insights

GitHub Actions workflows used to run automated tests, release Python packages on PyPI and push Docker images to Docker Hub. Dependabot for automatic updates of dependent Python packages.

GitHub project (02/2022-)

https://github.com/orgs/HEPData/projects



- HEPData/hepdata: 190 closed issues since 10/2020, but 208 created.

Modes of data entry

1. Manually harvested from data given in publications.
   HEPData staff extracted tables from `.tex` source.

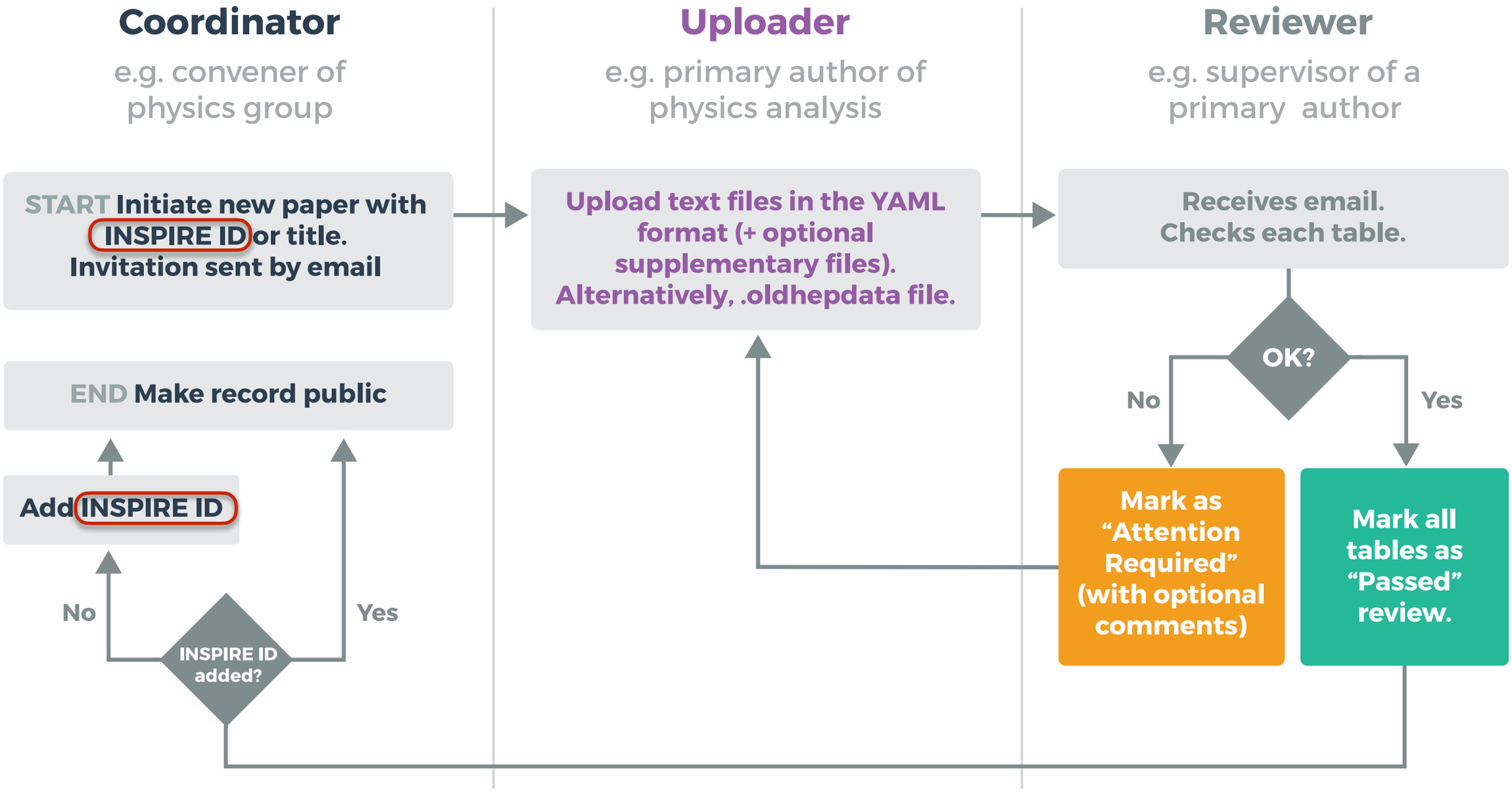
2. Data points directly submitted by experiments.

   - Pre-2014: no guidelines on preferred format.
   - Early 2014: encourage standard "input" format.
   - Late 2014: introduce online submission system.
   - Early 2017: allow submissions from hepdata.net.

Mode 1. now phased out in favour of mode 2.

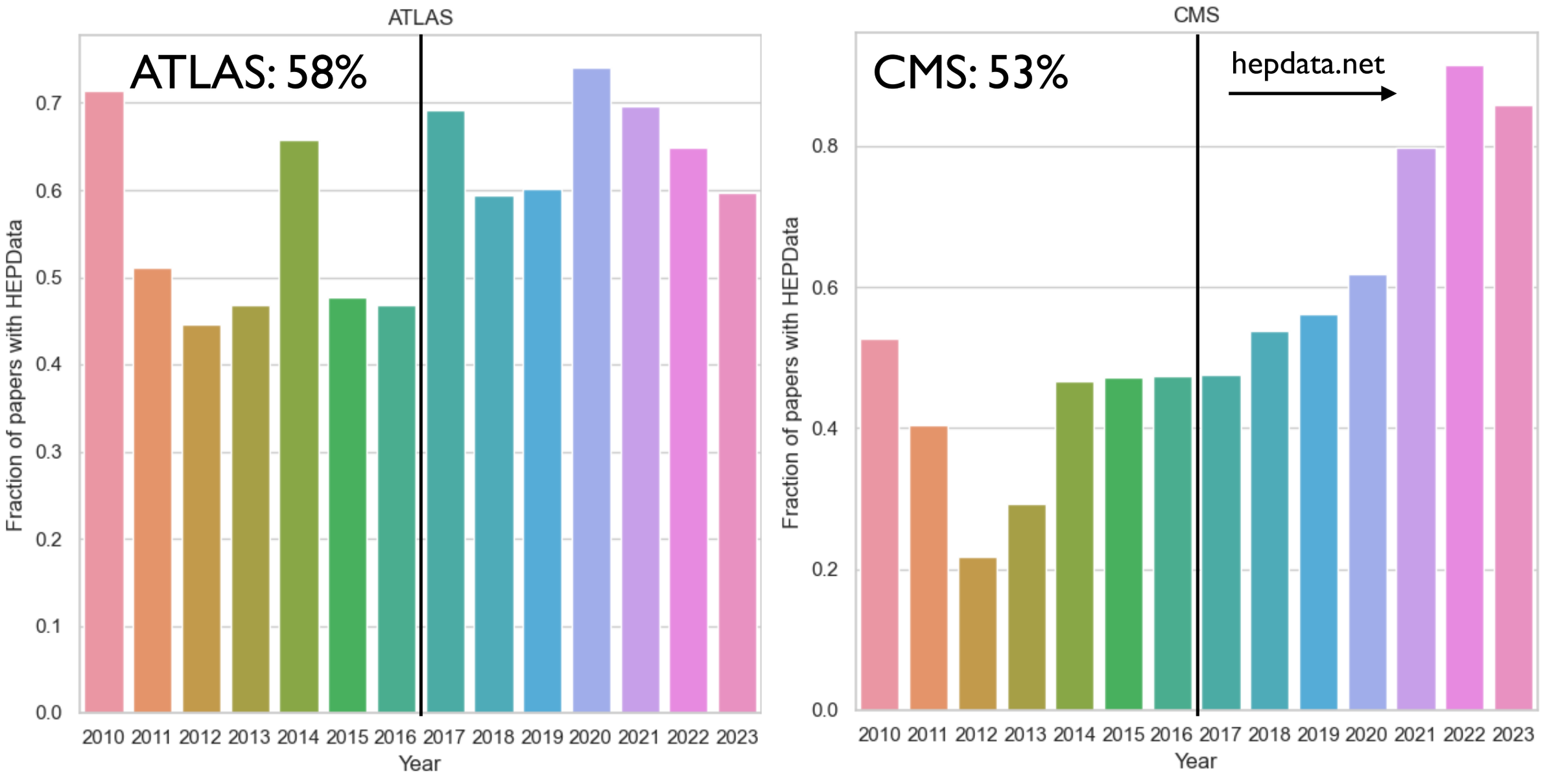
Submission system on hepdata.net

https://hepdata.net/submission

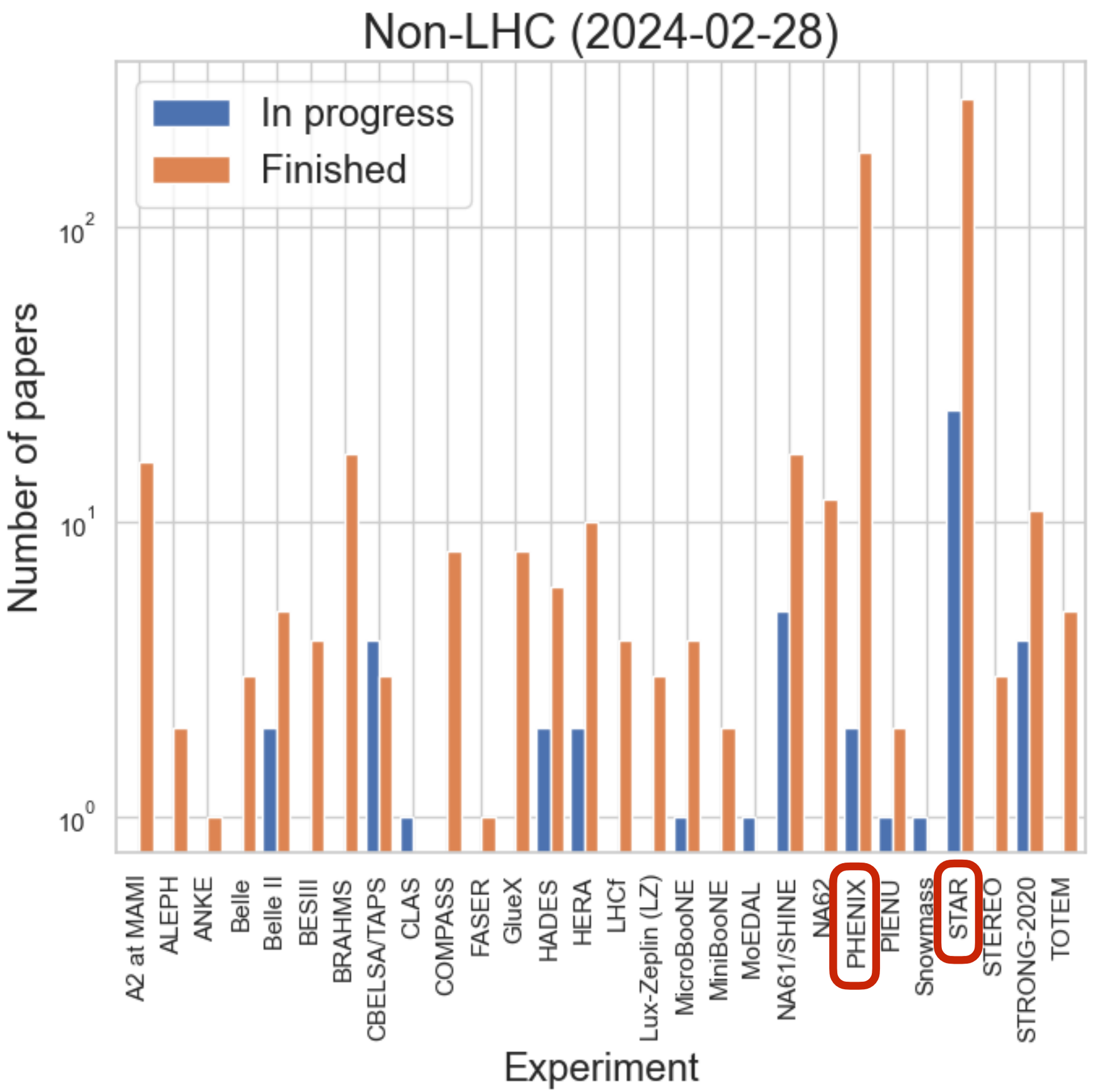


- Submissions managed by Coordinators within each experiment/group.

Coverage of ATLAS/CMS publications



- Search INSPIRE for publications with HEPData (GitHub/Binder).



- Big efforts by STAR and PHENIX at RHIC (BNL News).

https://www.bnl.gov/newsroom/news.php?a=221570

Physicists and Students Format PHENIX Data for Easy Access

Effort standardizes data needed to unlock the secrets of matter while building skills and bringing new faces into science

November 29, 2023

Christine Nattrass, a physics professor at the University of Tennessee (UT), Knoxville, has recruited a crew of mostly undergraduate students to dig deep into data from billions of particle collisions at the Relativistic Heavy Ion Collider (RHIC)—a U.S. Department of Energy (DOE) Office of Science user facility for nuclear physics research at DOE's Brookhaven National Laboratory. Their goal: reformat data from scientific papers published by RHIC's PHENIX detector collaboration and upload it to a modern database now used across the nuclear and high energy physics (HEP) research communities.

Posting the PHENIX data to this database, known as "HEPData," would make it accessible to anyone wanting to compare new findings with historical measurements or results from one experiment to another—or see how experimental results match up with theoretical descriptions of the building blocks of matter.

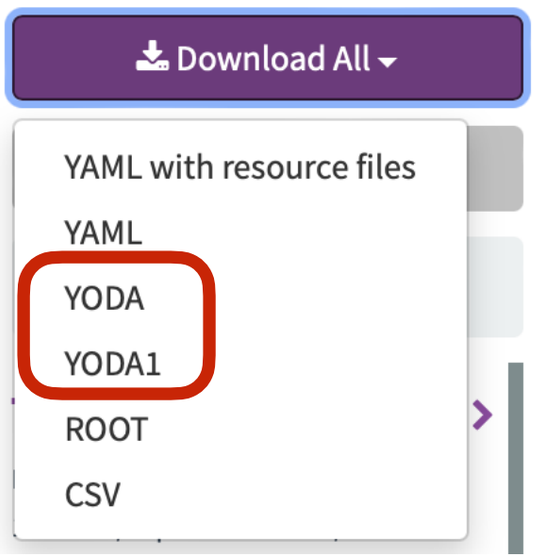
Data output formats

hepdata.net/formats

YAML: native HEPData format.

`submission.yaml`
+ YAML data files for each table
+ optional resource files

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



- NEW "YODA" now gives new YODA2 format.
- Legacy YODA format still available as YODA1.
- Thanks to Chris Gütschow (UCL) for work on implementing the YAML → YODA2 conversion.
- YODA2 publication: arXiv:2312.15070

Links to Rivet analysis code

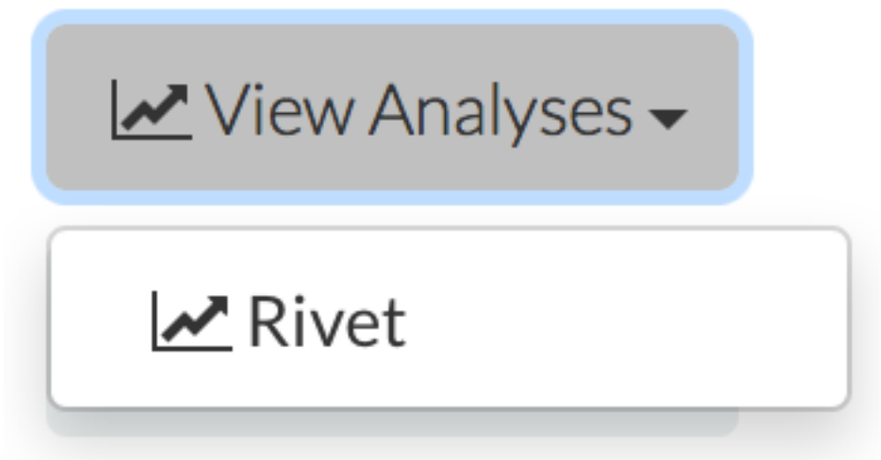
http://rivet.hepforge.org/analyses.json

- JSON file maps INSPIRE IDs to Rivet analysis names:

```
{"100016": [“GAMMAGAMMA_1975_I100016”], …,
     “954993”: [“ATLAS_2011_I954993”]}
```

- 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.

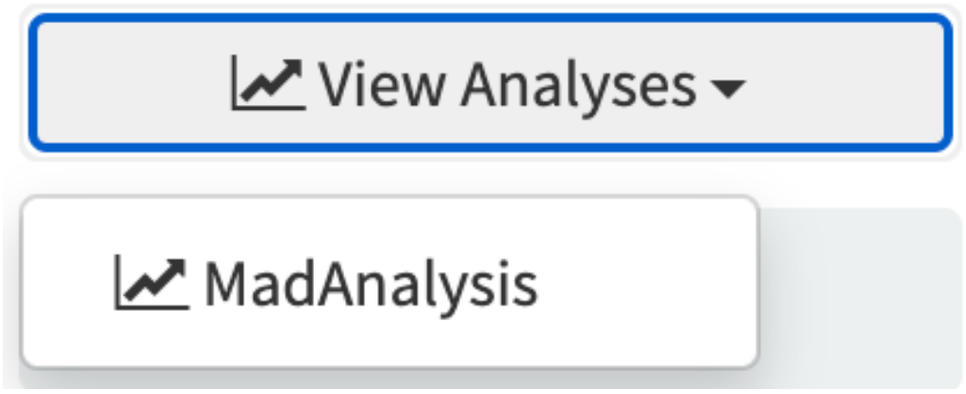
NEW Links to MadAnalysis 5

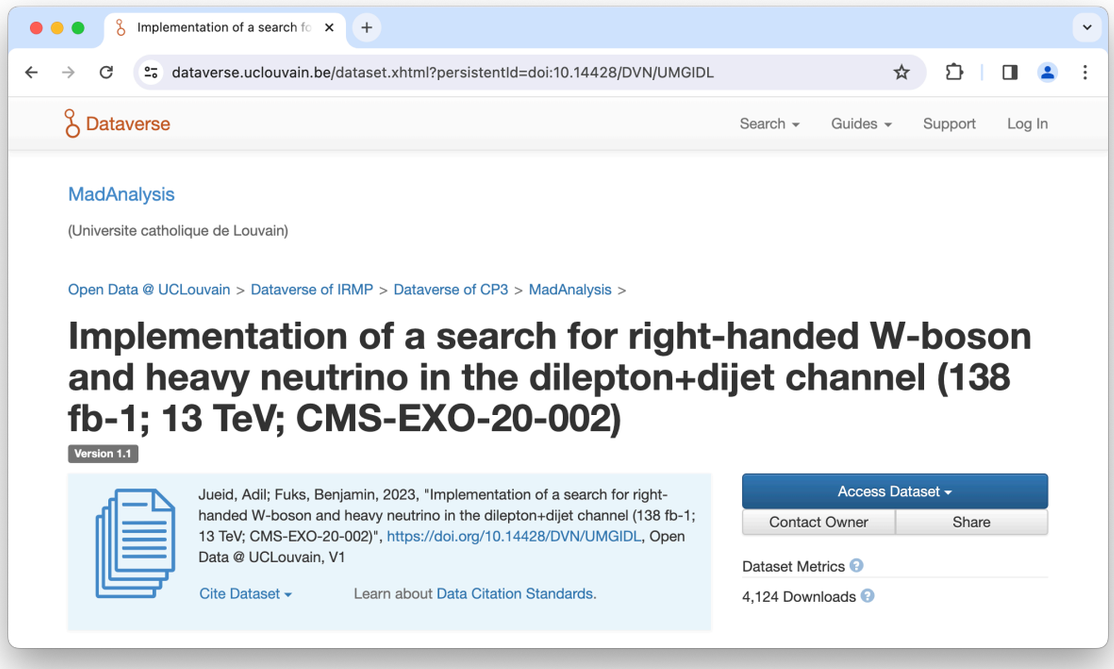
- analyses.json file for MadAnalysis 5 analyses:

```
{"1458270": [“10.14428/DVN/MHPXX4”], …,
  “1750186”: [“10.14428/DVN/OFAE1G"]}
```

- Search analysis:MadAnalysis gives 26 records.

MadAnalysis **Search for a right-handed W boson and a heavy neutrino in proton-proton collisions at $\sqrt{s}$ = 13 TeV**





- Thanks to Jack Araz.

Neutrino cross sections

- HEPData talk presented by G.W. at "IPPP/NuSTEC topical meeting on neutrino-nucleus scattering" back in April 2017.
- Attempts made to engage neutrino experiments to submit to HEPData, but limited uptake (e.g. MicroBooNE, MiniBooNE).
- Met with *Luke Pickering* and *Patrick Stowell* in May 2023.
- Plans to migrate NUISANCE data collection to HEPData and standardise future data releases from neutrino experiments.
- Talk by *Luke Pickering* on "Towards a Standardised Data Release Format" at "NuXTract 2023 - Towards a consensus in neutrino cross sections" CERN workshop in October 2023.

Submission documentation

- Documentation at hepdata-submission.readthedocs.io. Includes example Python scripts (simple, complicated).
- HEPData YAML files checked against JSON schema by validation code during submission. `hepdata-validate`
- hepdata_lib package by *Clemens Lange* (and *Andreas Albert*). Library to read in text/ROOT and write HEPData YAML. https://github.com/HEPData/hepdata_lib
  🆕 Convert from scikit-hep/hist histograms by *Yi-Mu Chen*.
- Similar Python package by *Christian Holm Christensen*. https://gitlab.com/cholmcc/hepdata
- Experiments often develop internal HEPData tools/docs.

hepdata-cli

- CLI and Python API for HEPData search/download/upload.
- Summer project in 2020 by Giuseppe De Laurentis.
- Install (in venv) with: `pip install hepdata-cli`
- Examples of usage:

```
hepdata-cli find 'collaborations:"Belle-II"' -i inspire
```

```
hepdata-cli fetch-names 1860766 -i inspire
```

```
hepdata-cli download 1860766 -f csv -i inspire
```

```
hepdata-cli upload /path/to/TestHEPSubmission.tar.gz -e
my@email.com -p $PASSWORD –r 123456 -i $INVITATION_COOKIE -s False
```

Code: https://github.com/HEPData/hepdata-cli

NEW Bidirectional linking

- Suggestion by Jon Butterworth in December 2022. Technical implementation by Jordan Byers (Durham).
- Enable bidirectional links *between* HEPData **tables** possibly in different records in `submission.yaml`:

```yaml
related_to_table_dois:
- 10.17182/hepdata.12345.v1/t2
- 10.17182/hepdata.67890.v3/t4
```

or use hepdata_lib

- Similar bidirectional links *between* HEPData **records**:

```yaml
related_to_hepdata_records:
- 12345
- 67890
```

or use hepdata_lib

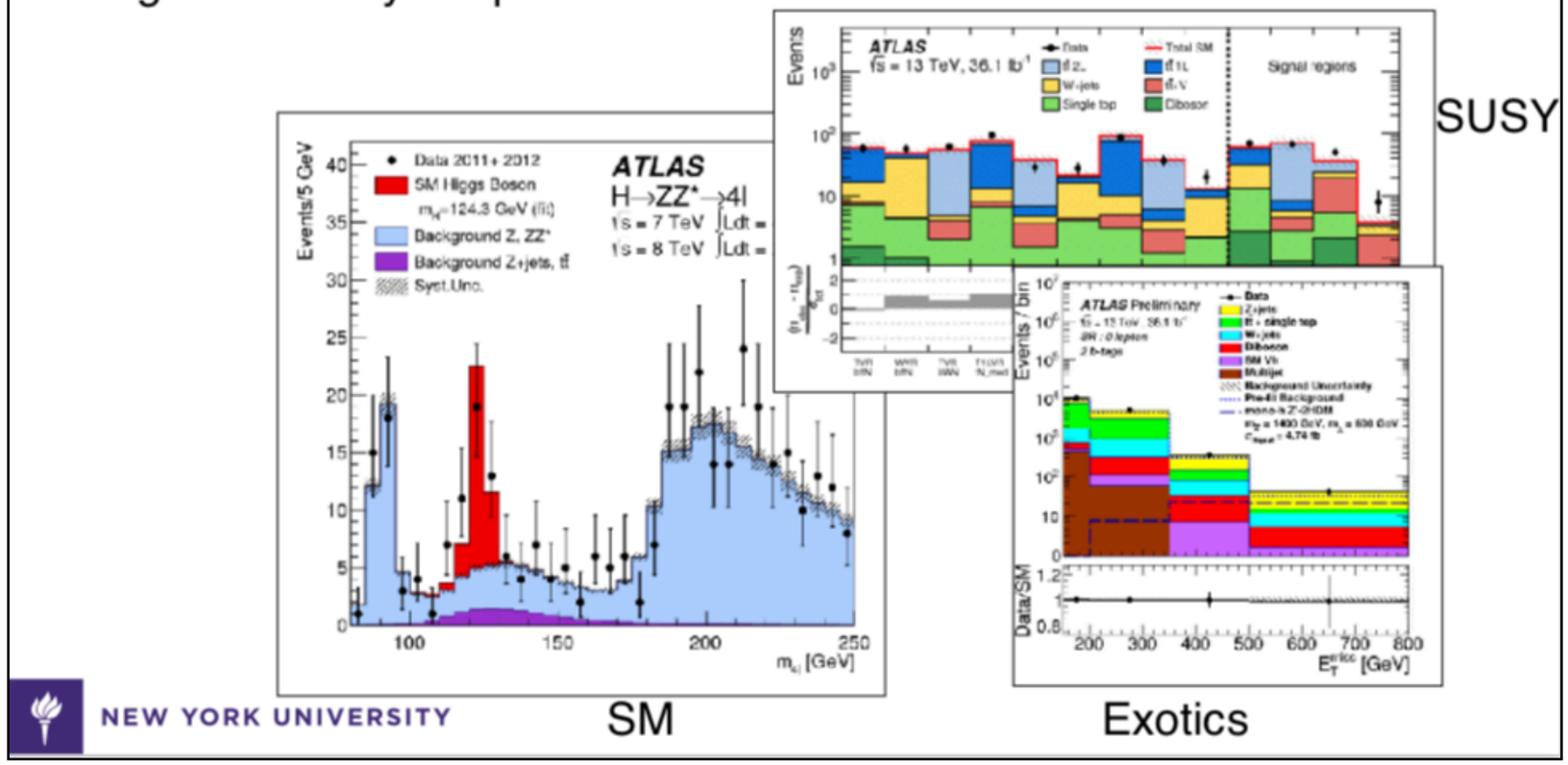
HistFactory

HistFactory

Lukas Heinrich

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

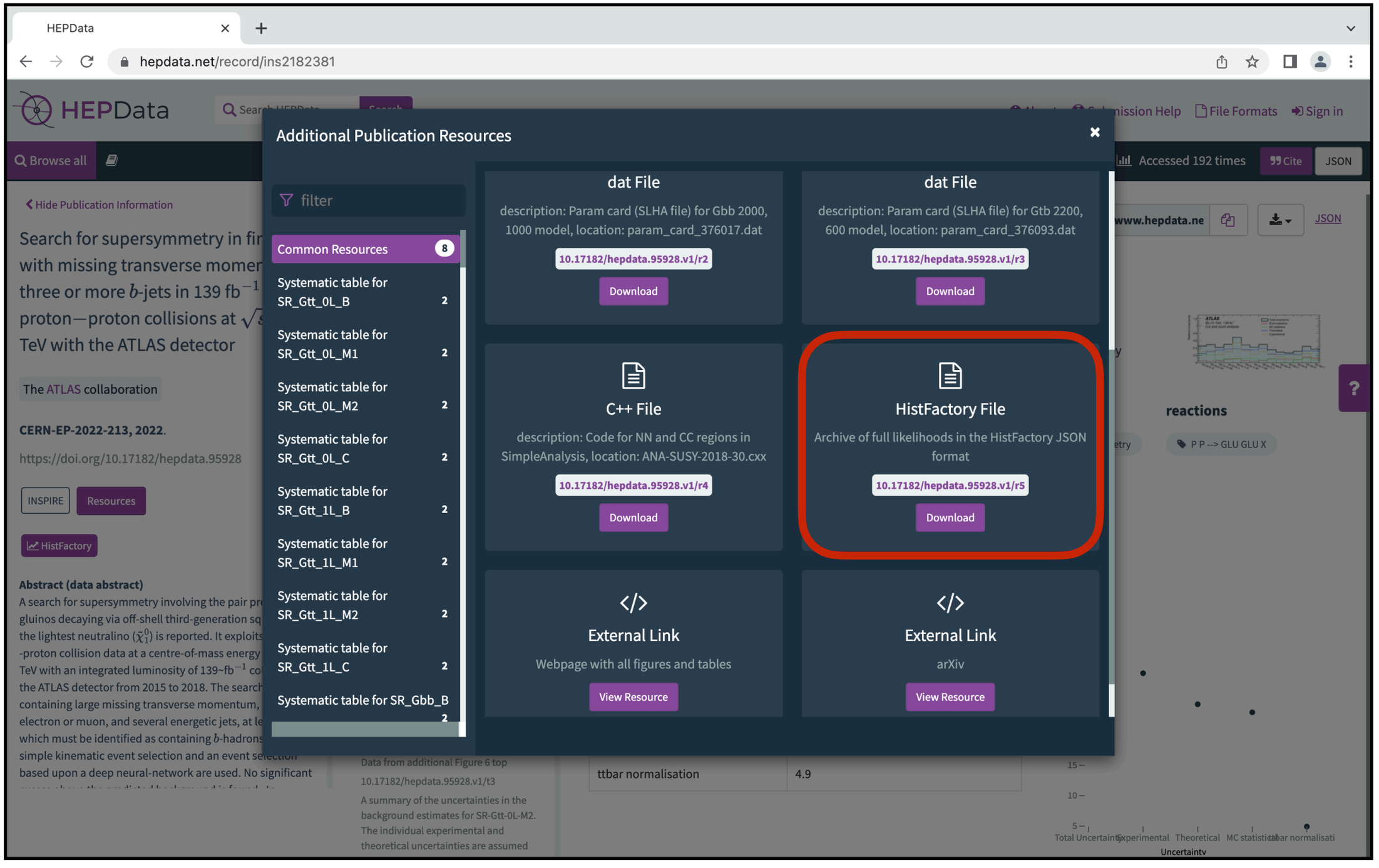
~all binned, folded, frequentist statistical models in ATLAS are expressed using HistFactory template



NEW YORK UNIVERSITY

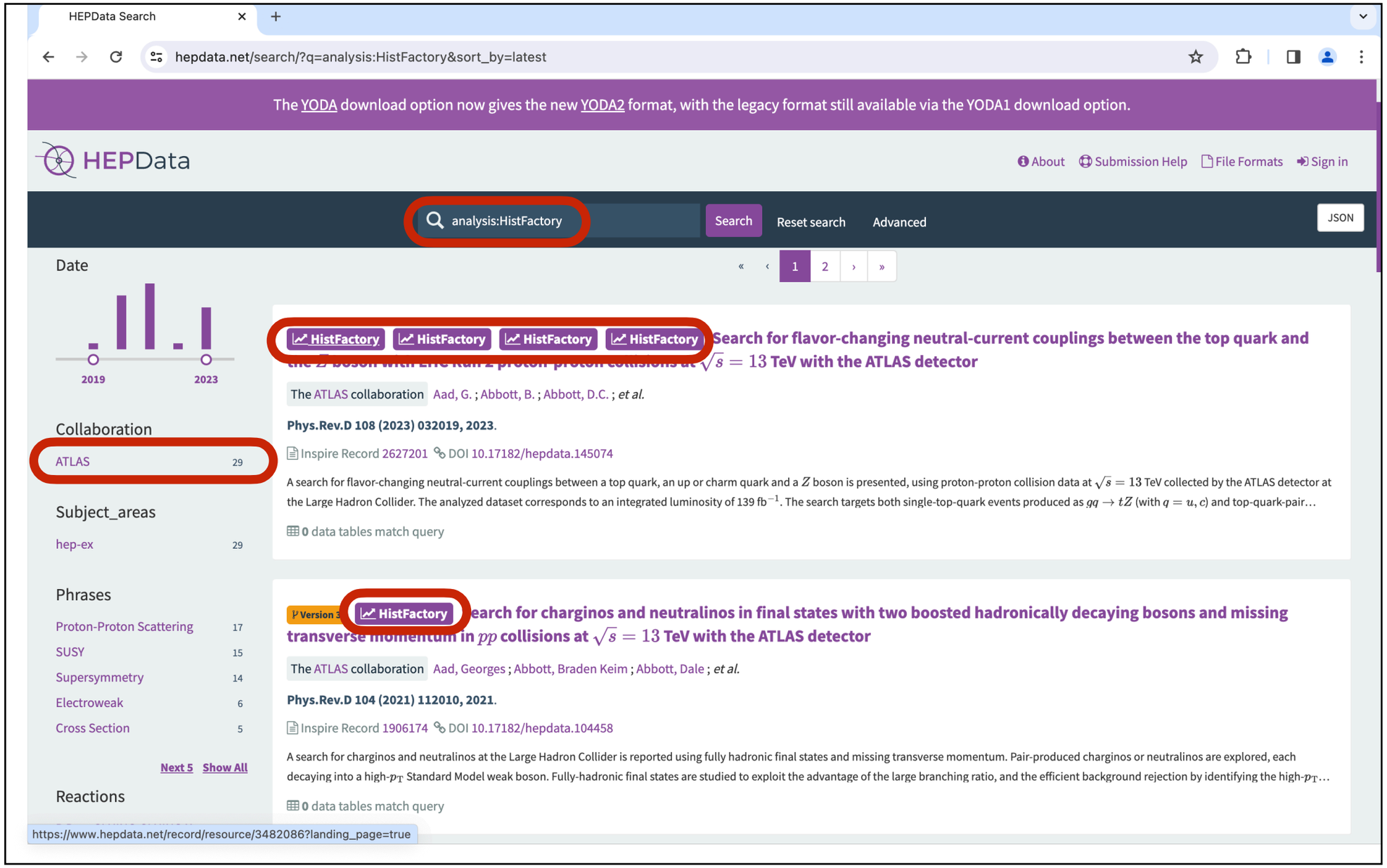
- Based on simple ROOT histograms organised in an XML file.
- ROOT/XML workspace replaced by plain-text pyhf JSON.

Additional resource files



- DOIs now minted via DataCite for additional resource files.

pyhf likelihoods



- Search `analysis:HistFactory`.

Native support for pyhf JSON?

- **Idea:** support submissions with pyhf JSON files replacing the usual HEPData YAML data tables. Provide a visualization of the pyhf JSON files within the HEPData record.

- First step completed by Sinclert Pérez (NYU) in 2020: modify `hepdata-validator` to use a remote JSON schema.

- Idea abandoned as too complicated and not well-defined.

- New idea (2022): standalone web application to visualize pyhf JSON data that can import resource files from HEPData.

- Two applications presented at pyhf workshop in Dec 2023:

1. HFExplorer (*Abe Megahed*): `hfexplorer.net`

2. WorkspaceExplorer (*Volker Austrup*): `workspaceexplorer.app.cern.ch`

OpenMAPP (03/2024 - 02/2026)

Open meta-analysis in particle physics

- Project funded by CHIST-ERA in call for "Open & Re-usable Research Data & Software".

Partners' people involved in the realisation of the project

Partner NB	Funder (if any)	Country	Institution / Department	Name of the Principal Investigator (PI)	Name of the co-Investigators	Name of the other personnel participating in the project
1 Coord.	ANR	France	LPSC Grenoble	Sabine Kraml		Post-doc
2	TUBITAK	Turkey	Bogazici University	Erkcan Ozcan		Gokhan Unel (advisor) 2 x post-doc 2 x PhD 2 x MSc
3	NCN	Poland	Jagiellonian University	Andrzej Siodmok		Post-doc
4	UKRI	UK	University of Glasgow	Andy Buckley		Post-doc
5	-	France	LPTHE	Benjamin Fuks	Mark Goodsell	
6	-	UK	University College London	Jonathan Butterworth		Christian Gutschow
7	-	UK	Durham University	Graeme Watt		
8	-	Korea	Kyungpook National U.	Sezen Sekmen		Junghyun Lee (PhD)

No	Task name	Lead	Partners	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1.1	Extension of HEPData functionalities	7	4,5,7	x	x	x	x	x	x																		
1.2	Interfacing to HEPData	4	1,2,4,5,6,7	x	x	x	x	x	x	x	x	x	x	x	x												
1.3	Interfacing to LHC Open Event Data	8	2,3,8							x	x	x	x	x	x												
2.1	(Meta)Database of implemented analyses	2	1,2,4,5,6,8			x	x	x	x	x	x	x	x	x	x	x	x	x									
2.2	Common interface for analysis steering and output	4	2,4,5,6,7			x	x	x	x	x	x	x	x	x	x												
2.3	MC interface and common validation infrastructure for recasting frameworks	3	2,3,4,5,6,8													x	x	x	x	x	x	x	x				
3.1	Statistical models	1	1,2,4-8								x	x	x	x	x	x	x	x	x								
3.2	Enable combinations of analyses	2	1,2,4,5,6,8												x	x	x	x	x	x	x	x	x	x			
3.3	Physics case study	1	1-6,8																					x	x	x	x

Follow @HEPData

Summary

Email: info@hepdata.net
Forum: hepdata-forum.cern.ch

- **HEPData** is *the* repository for publication-related HEP data.
- *Caveats*: design restricts size (~MB) and format (mostly tabular).
- Widely used by HEP community: 4.5 million page views in 2023.
- Infrastructure via CERN, development and support via Durham.
- Essential contributions from funded *Software Engineer* position.
- Open development process via https://github.com/HEPData .
- Enhanced linking to Rivet analyses and MadAnalysis 5 analyses.
- pyhf JSON likelihood data hosted as additional resource files.
- OpenMAPP to further enhance interfaces with analysis toolkits.