HEPData

and OpenMAPP

Graeme Watt (IPPP Durham)

Reinterpretation & OpenMAPP mini-workshop, Grenoble
Monday 17th June 2024

https://hepdata.net

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

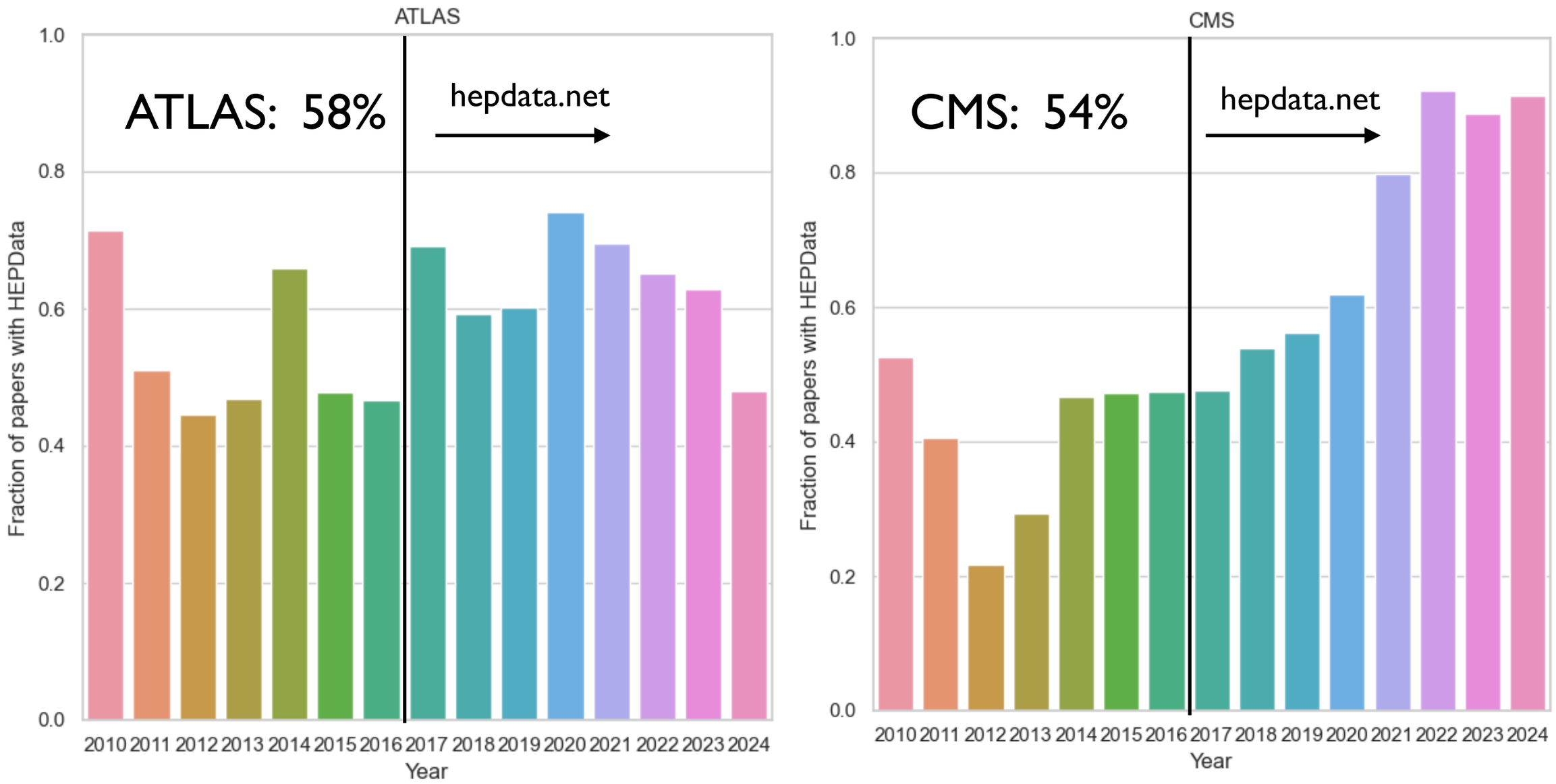
Follow @HEPData

Code: https://github.com/HEPData

What is HEPData?

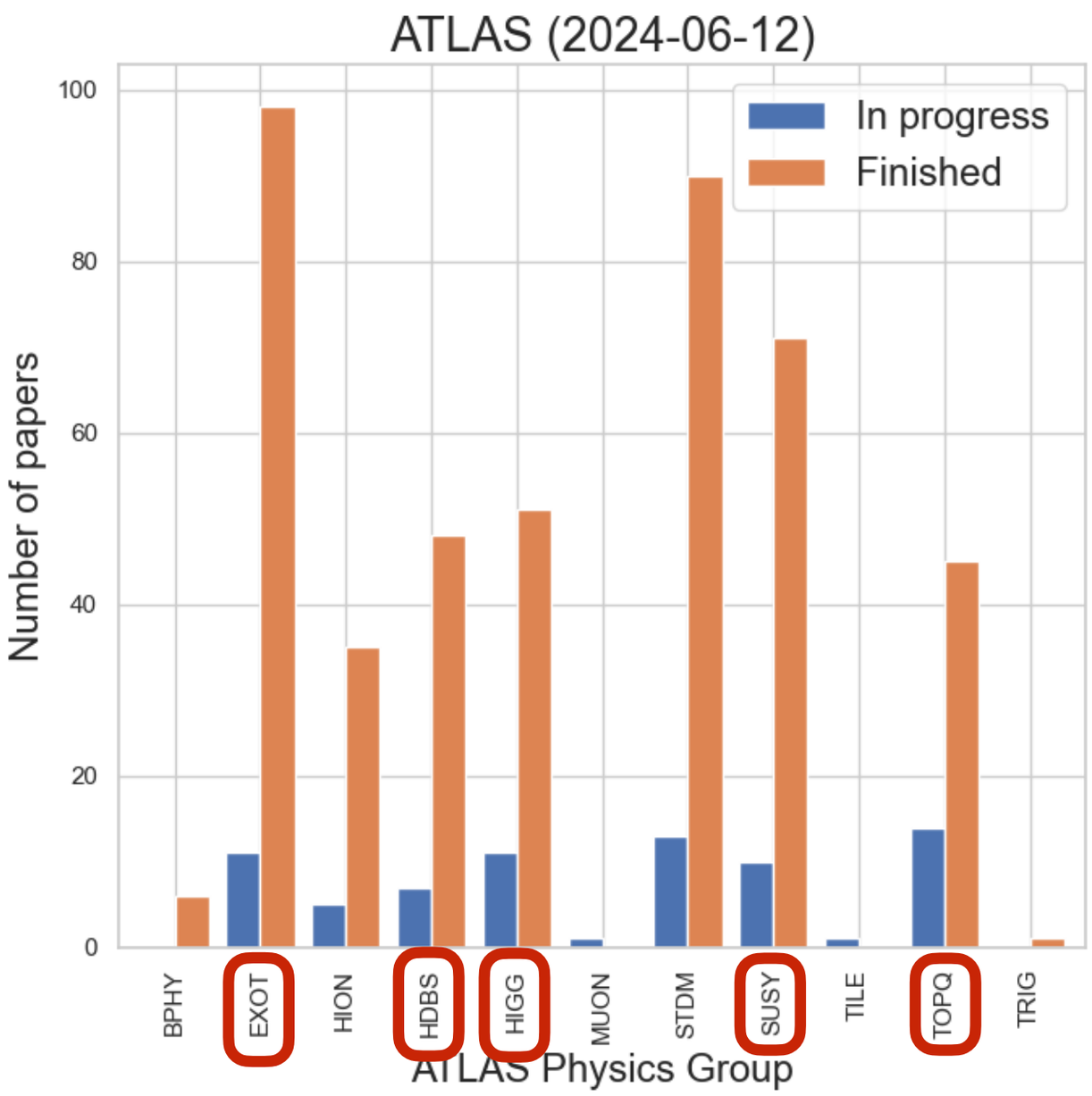
- Unique *open-access* repository for tabular high-level **data** from more than 10k **HEP** publications (130k data tables).
- **FAIR** data: **F**indable, **A**ccessible, **I**nteroperable, **R**eusable.
- *Complementary* to other HEP information providers, e.g. INSPIRE-HEP (literature), PDG (particle 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]
- *Staff*: G.W. (Manager, 2013-), Jordan Byers (RSE, 2022-).

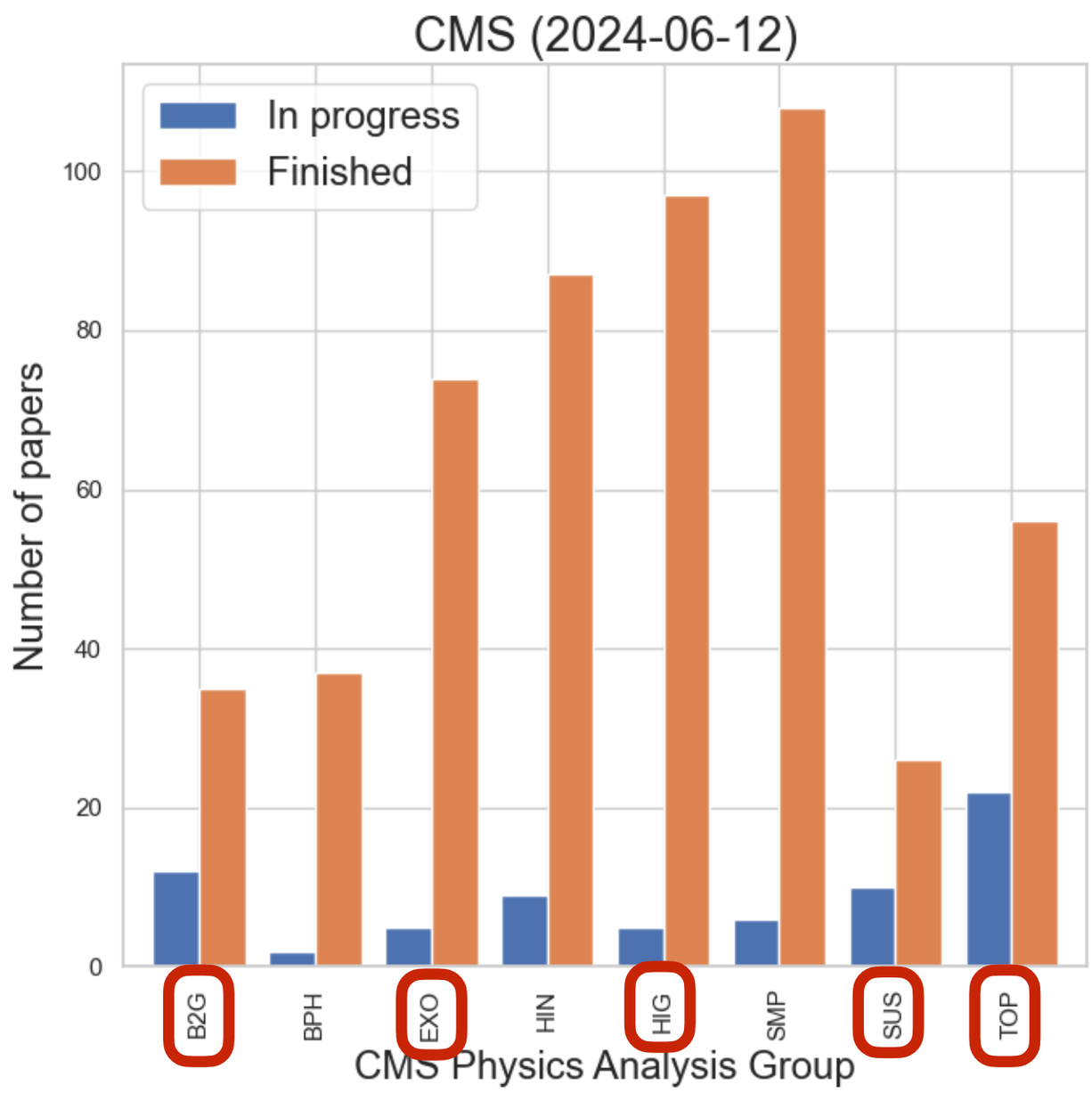
Coverage of ATLAS/CMS publications



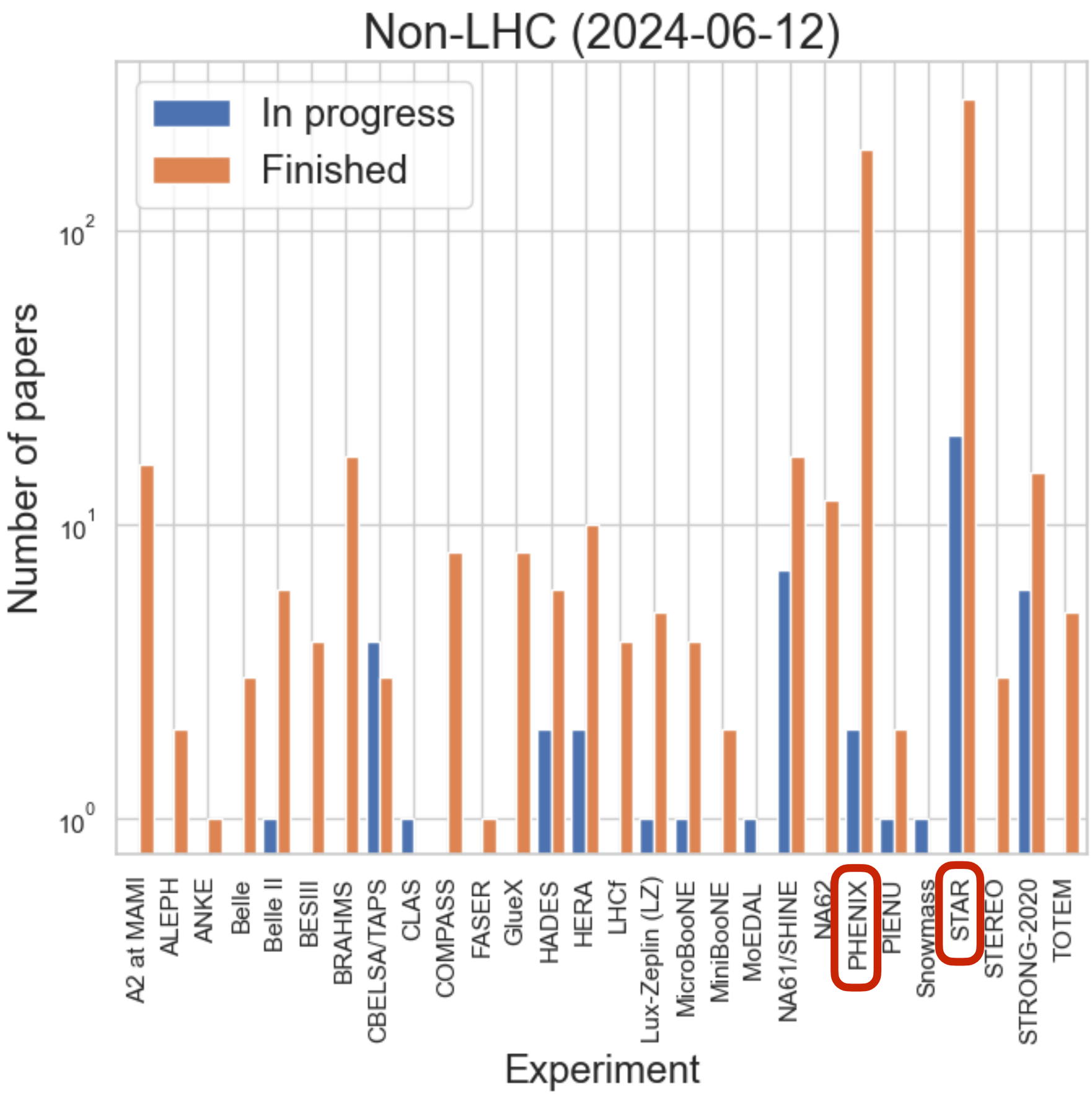
- Search INSPIRE for publications with HEPData (GitHub/Binder).

Submissions by ATLAS/CMS groups (2017-)



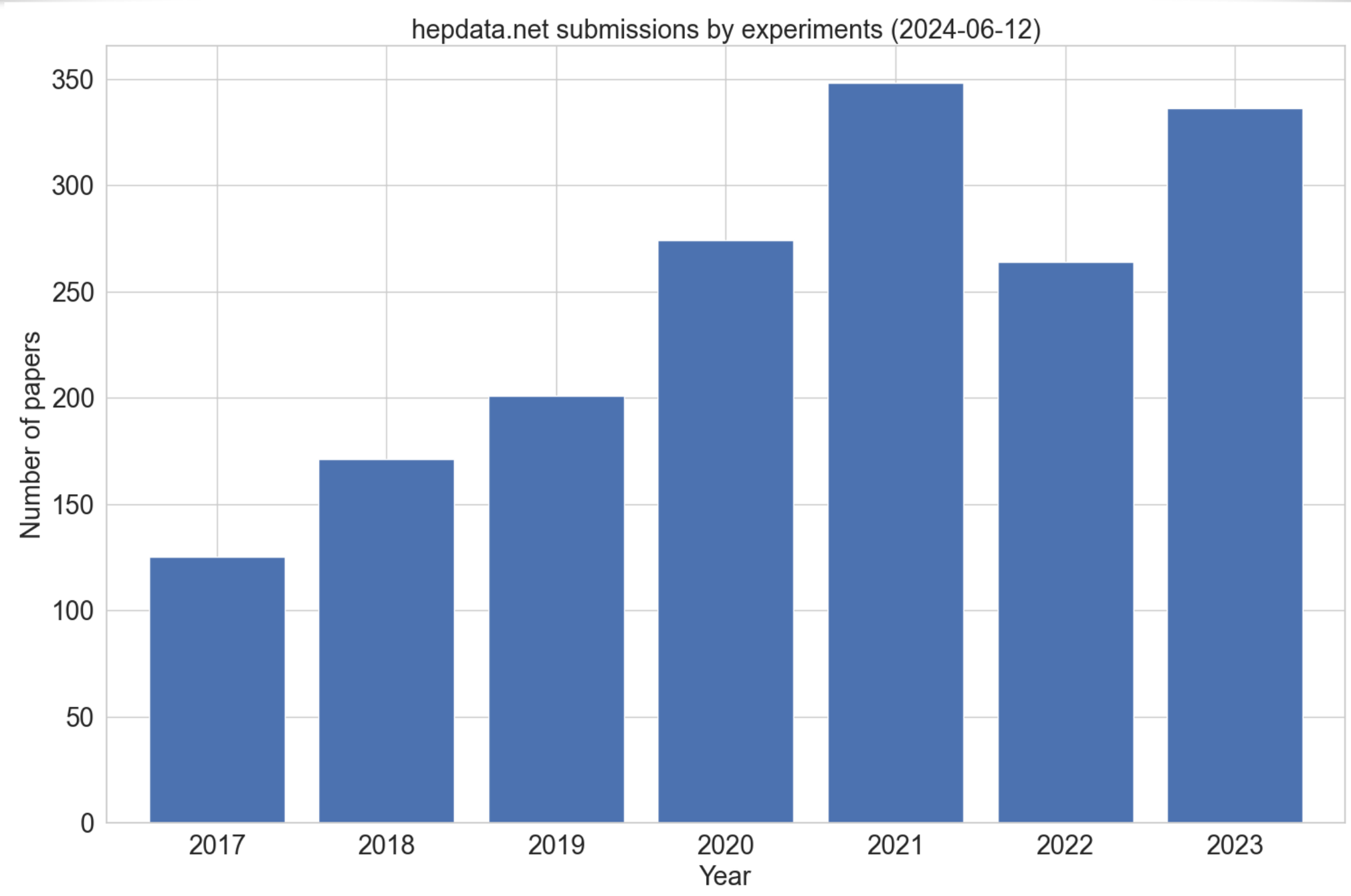


- Some ATLAS/CMS physics groups more active than others.



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

Submissions per year from 2017 to 2023



- Increase for first few years, then around 300 per year from 2020.

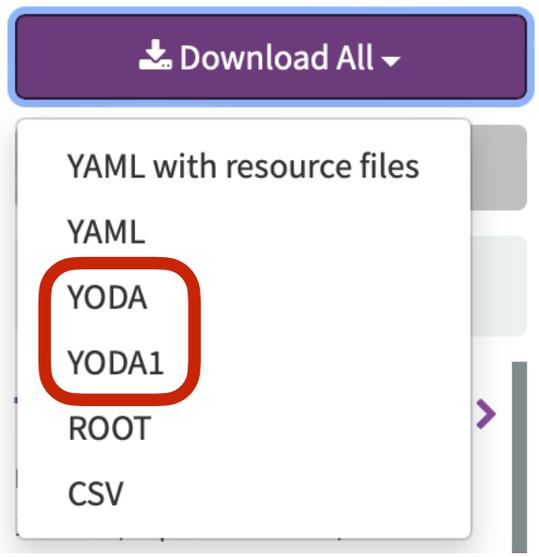
Data output formats

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.



from 17th Nov 2023

- NEW "YODA" now gives new YODA2 format, for use with Rivet version 4 (released Feb 2024).
- Legacy YODA format still available as YODA1.
- Thanks to Chris Gütschow (UCL) for work on implementing the YAML → YODA2 conversion.

Links to Rivet analysis code

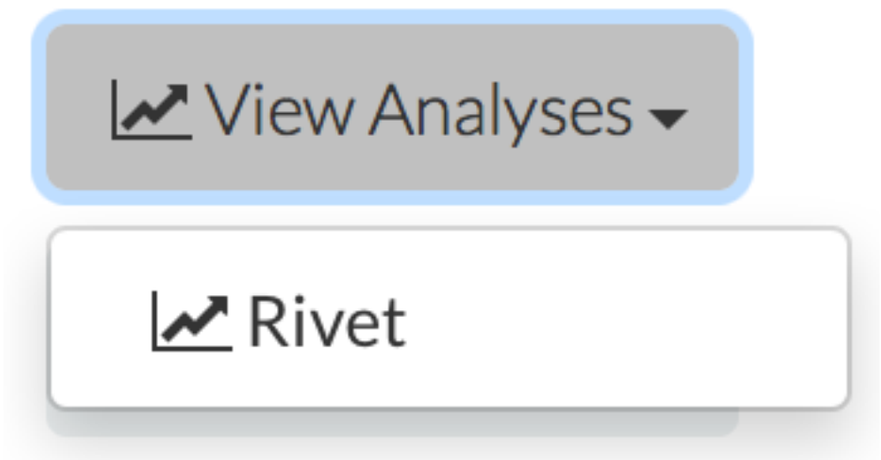
http://rivet.hepforge.org/analyses.json

- 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

from 5th Oct 2023

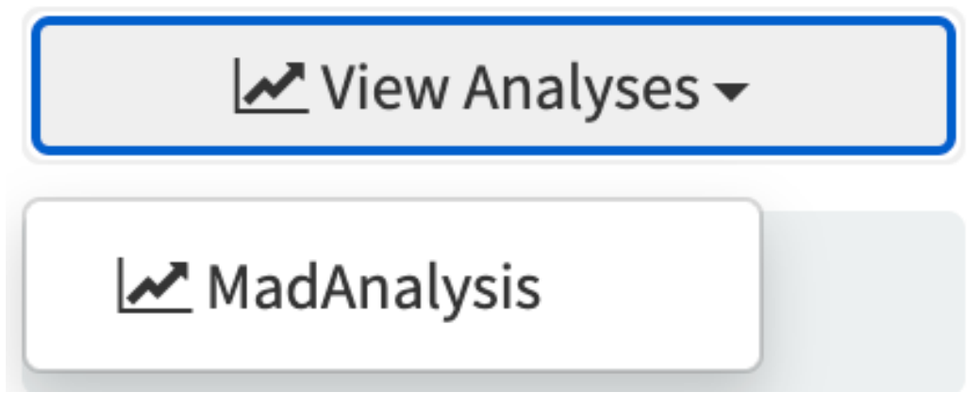
Links to MadAnalysis 5

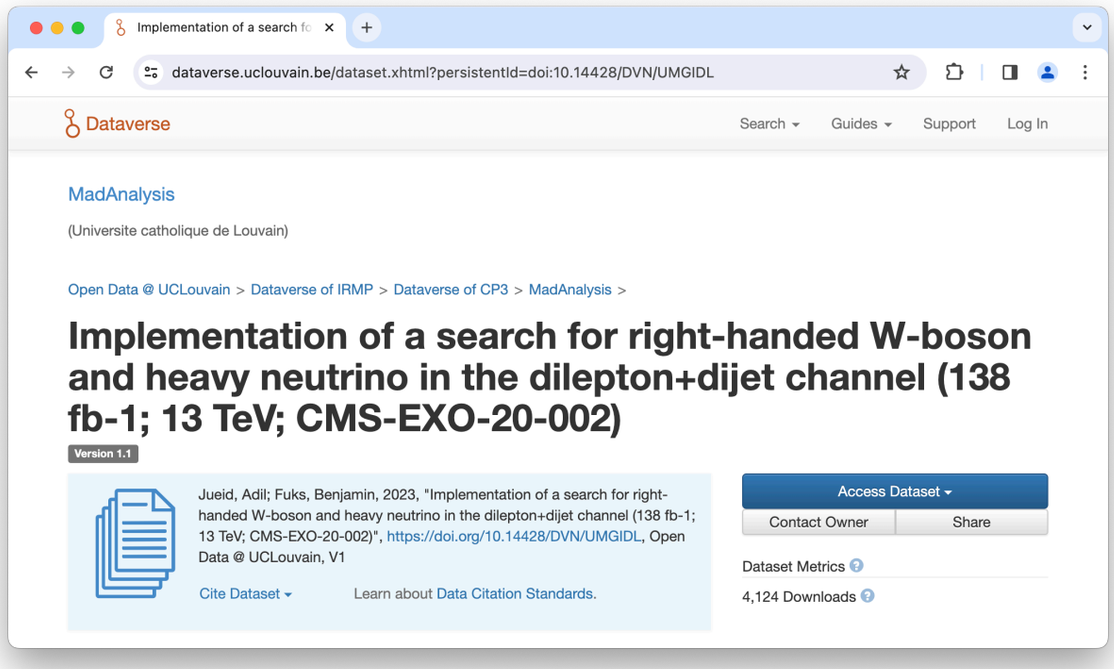
- `analyses.json` file for MadAnalysis 5 analyses:

```
{"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.

hepdata_lib

- 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

- **Recent improvements** by new contributors (releases):
  - Convert from scikit-hep/hist histograms by *Yi-Mu Chen*.
  - Revamp of CI workflow by *Matthew Feickert* (and *G.W.*).
  - Update docs for add_additional_resource by *G.W.*
  - Fix broken code documentation on *Read the Docs* by *G.W.*
  - Add option to suppress warning if zero uncertainties by *G.W.*
  - Allow some functionality without ROOT installation by *G.W.*
  - Remove pin on Pylint to support Python 3.11 by *G.W.*
  - Add functions for related records and tables by *Jordan Byers*.
  - Add functions to provide license information by *Jordan Byers*.

- **Future:** add ability to read in YODA files (issue)?

NEW

from 31st May 2024

Licenses

- HEPData Terms of Use:
  Unless specified otherwise for selected datasets, all metadata and datasets in the HEPData service are made available under the terms of CC0.

> CC0 (aka CC Zero) is a public dedication tool, which enables creators to give up their copyright and put their works into the worldwide public domain. CC0 enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, with no conditions.

- Specify a license other than CC0 using fields `data_license` (for data tables) or `license` (for resources) in `submission.yaml` file.

- Recent changes (motivated by feedback from CMS & Lukas Heinrich):
  - Render license (including default CC0) on HEPData web pages.
  - Functions to add license information included in `hepdata_lib`.

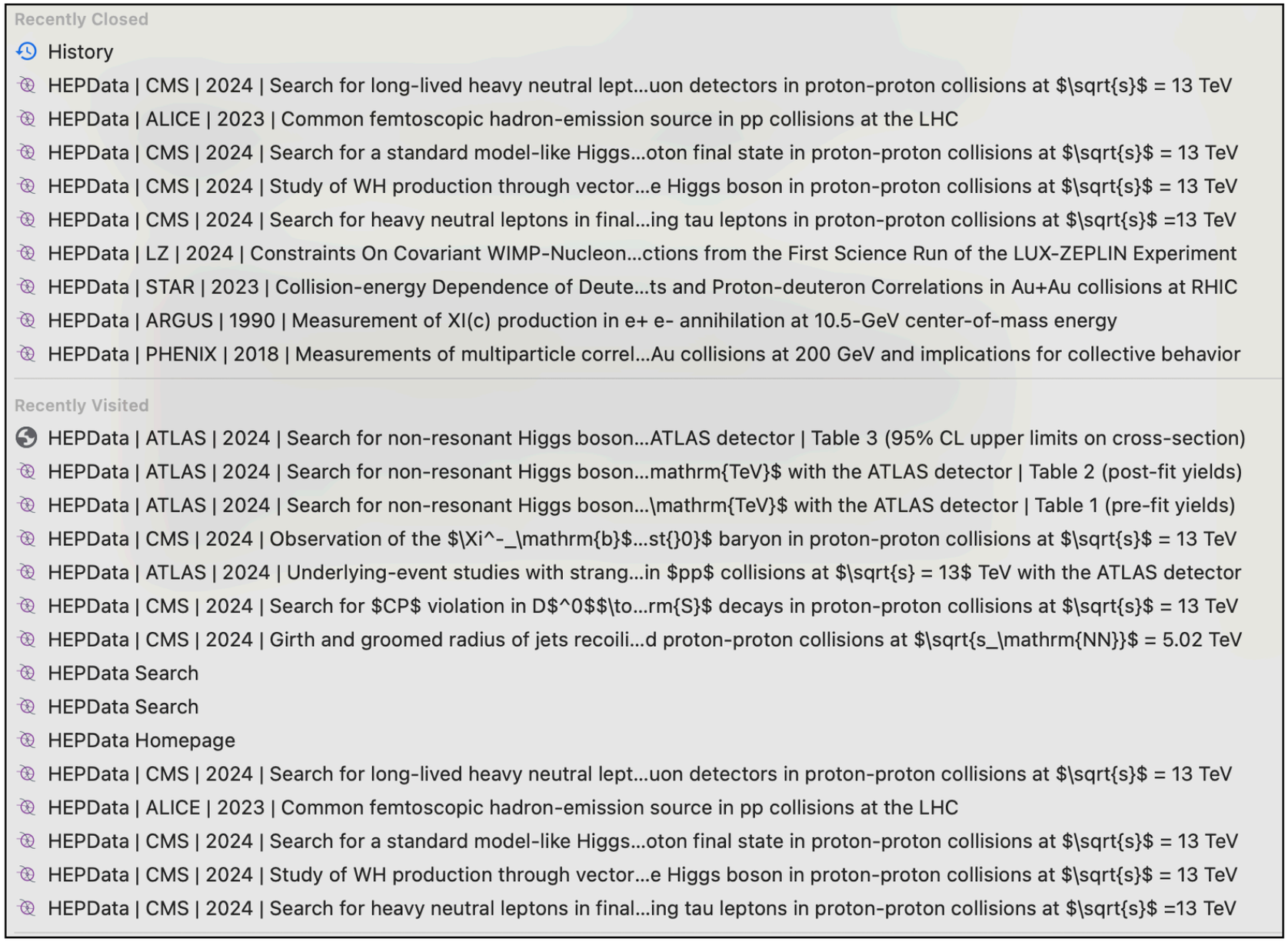
- CMS Higgs boson observation statistical model uses CC BY 4.0.

NEW

from 22nd Aug 2023

Record-specific <title> tags

- Suggestion by *Andy Buckley* in August 2023 for easier management of a user's browser history.

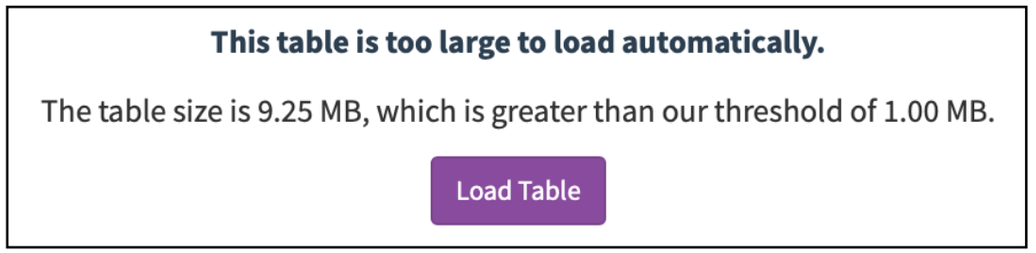


NEW

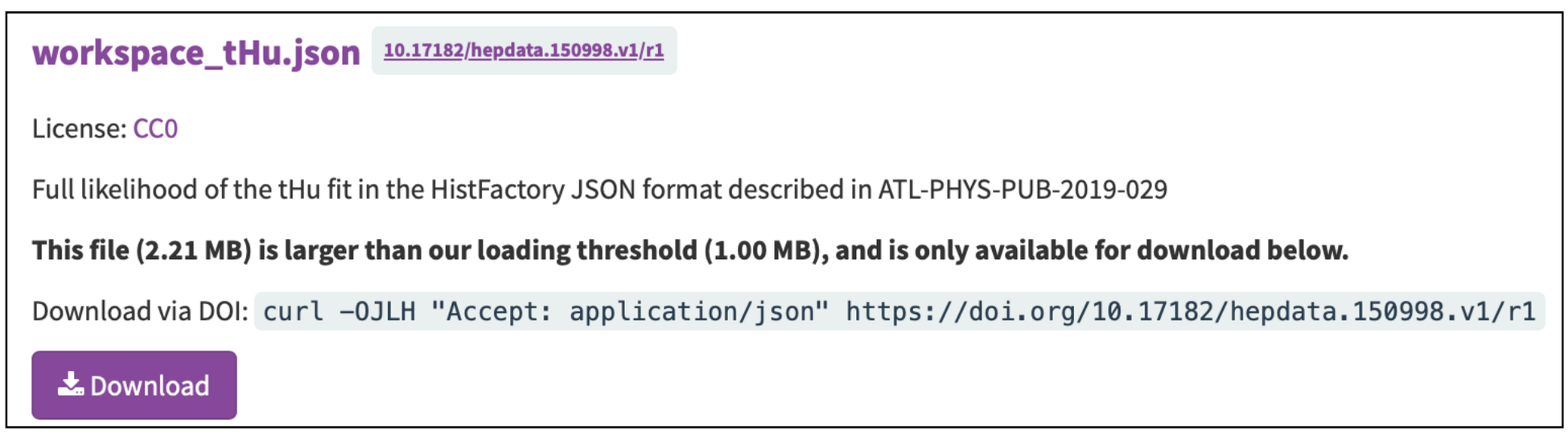
from 7th Dec 2023

Deferred loading of large files

- Validator restricts YAML data files to be smaller than 10 MB.
- YAML data files 1-10 MB can be slow to load in browser.
- Introduce deferred table loading if YAML data file > 1 MB:



- On resource file landing page if text file > 1 MB, only display "Download" button without trying to display in browser:



NEW

from 22nd Aug 2023

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**:

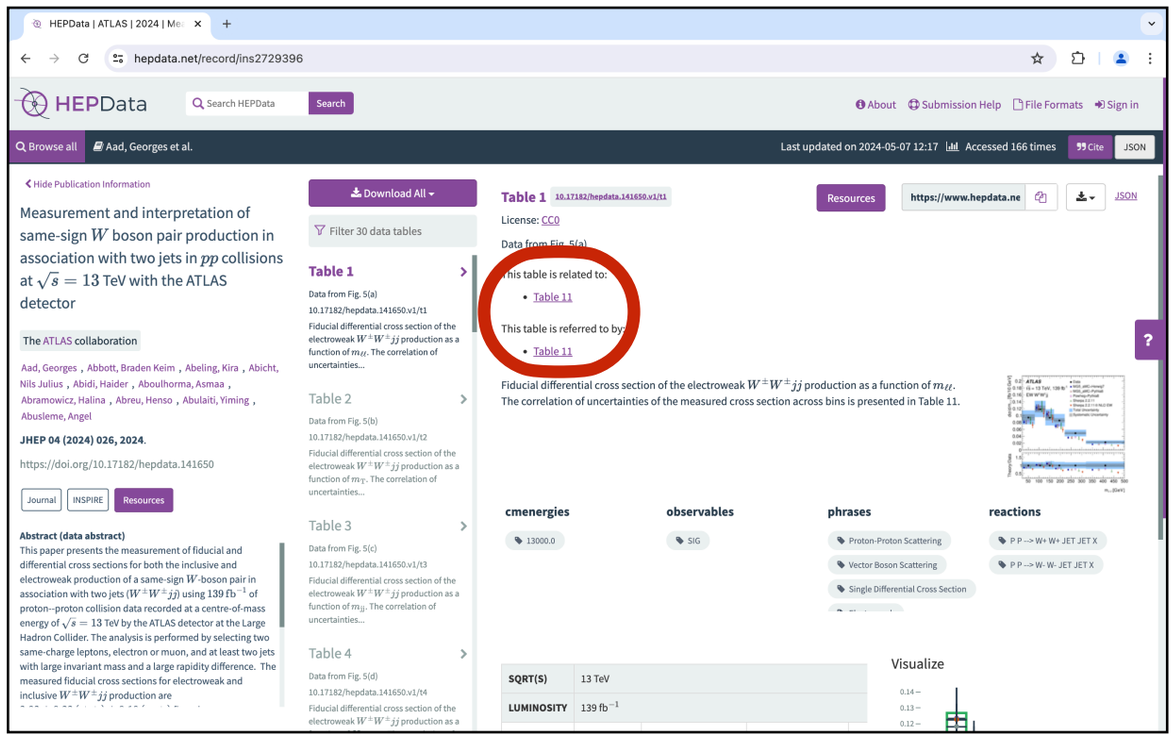
```yaml
related_to_hepdata_records:
- 12345
- 67890
```

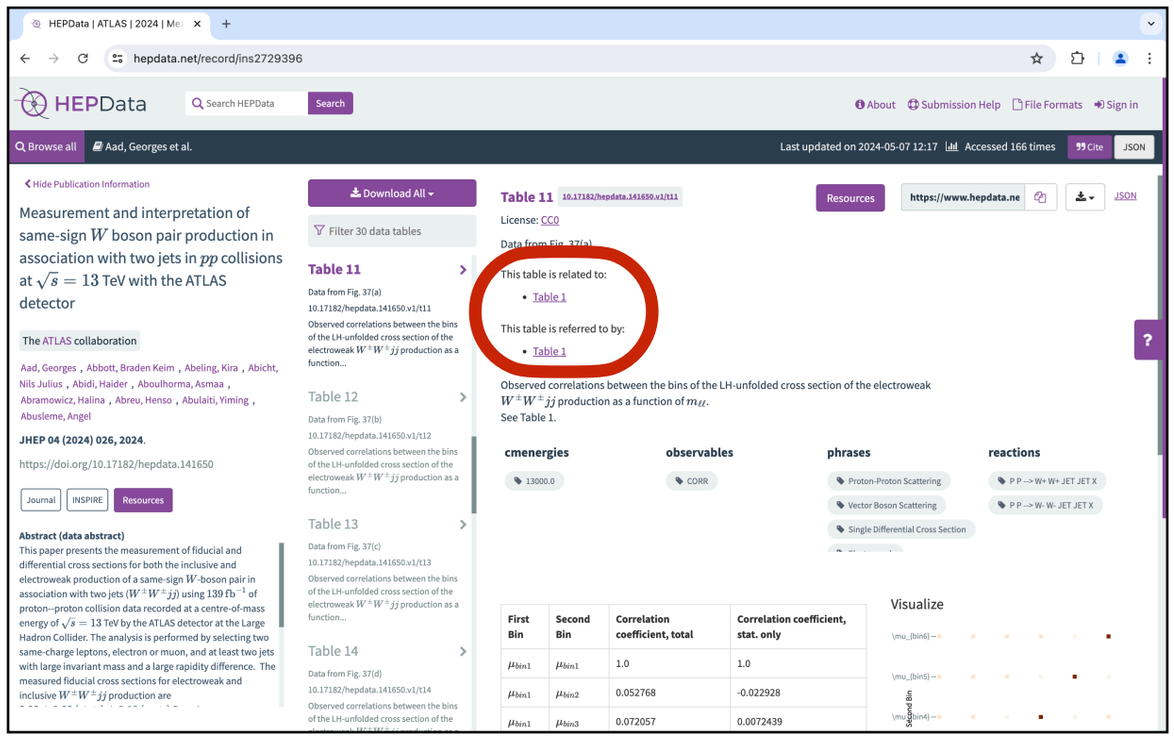
or use hepdata_lib

NEW

from 22nd Aug 2023

Bidirectional linking





`related_to_table_dois: [10.17182/hepdata.141650.v1/t11]`

~~`related_to_table_dois: [10.17182/hepdata.141650.v1/t1]`~~

- First record: https://www.hepdata.net/record/ins2729396
- Links are automatically *bidirectional*: only specify one direction.
- Issue with previewing bidirectional links before record finalised.

NEW

from 3rd May 2024

Searching for resources

Searching resources by field

Text-based description searching:

resources:"Created with hepdata_lib"

Quotes force a full match.

Resource-type searching:

resources.type:png

Examples: png, html, github, zenodo etc.

Searching for specific URLs:

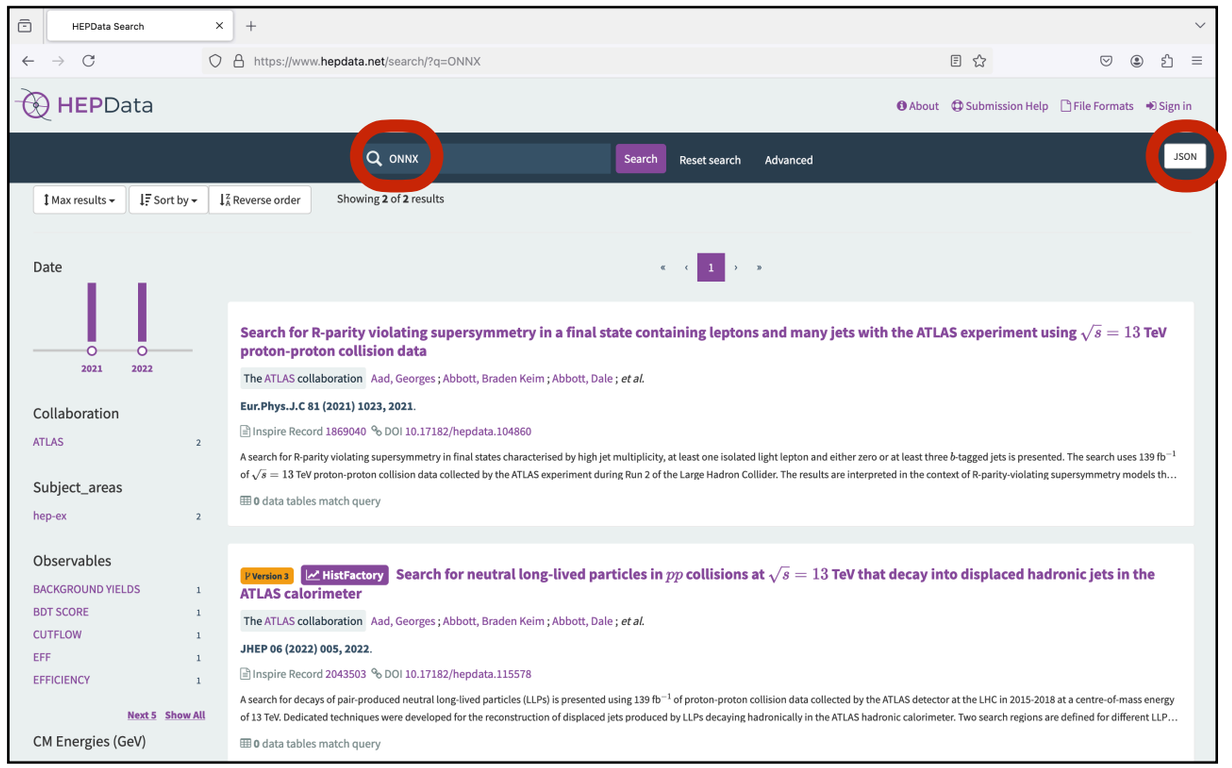
resources.url:atlas.web.cern.ch

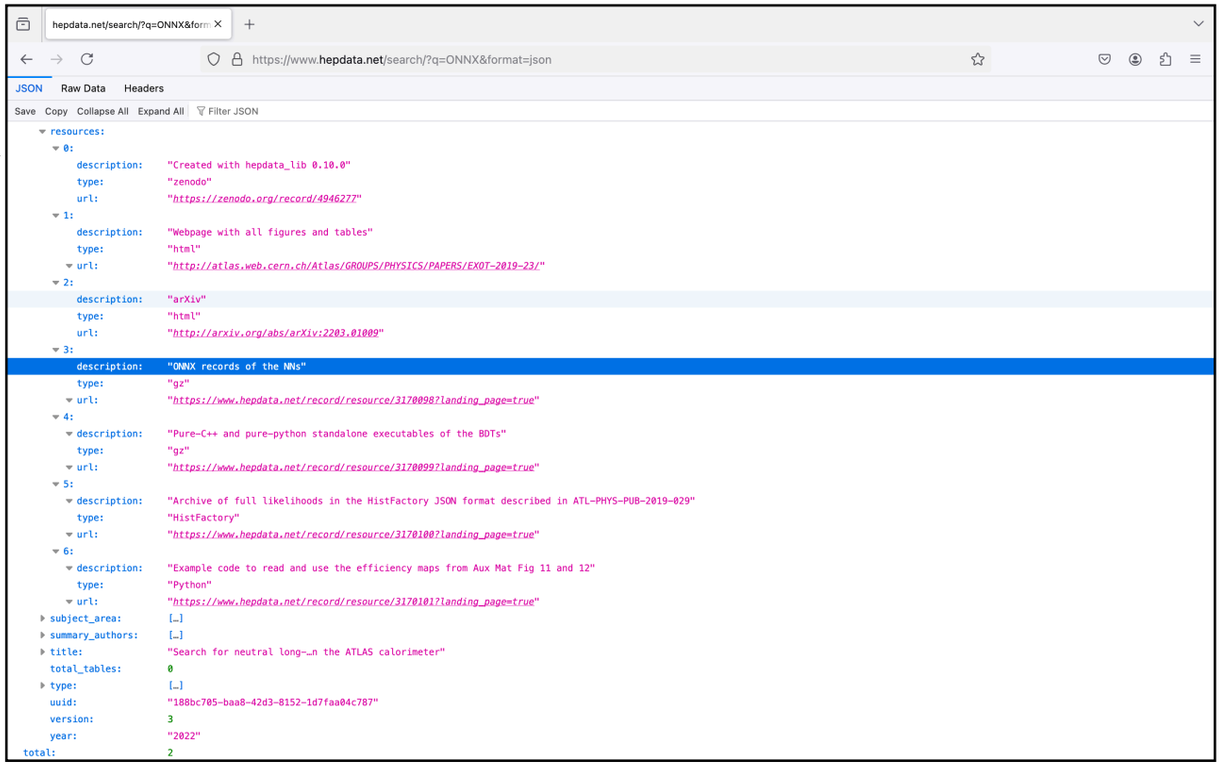
- Additional resource metadata now indexed for searching.

NEW

from 3rd May 2024

Resource information in JSON






```python
import requests
query = 'ONNX'
url = f'https://www.hepdata.net/search/?q={query}&format=json'
request = requests.get(url).json()
for result in request['results']:
    for resource in result['resources']:
        if query in resource['description']:
            print(result['inspire_id'])
            print(resource['description'])
            print(resource['type'])
            print(resource['url'].replace('landing_page', 'view'))
            print()
```

```
1869040
ONNX files for the neural networks for the EWK analysis
tgz
https://www.hepdata.net/record/resource/2677521?view=true

2043503
ONNX records of the NNs
gz
https://www.hepdata.net/record/resource/3170098?view=true
```

- **Example:** get download links of ONNX files from Python.

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`

- Reproduce previous example using `hepdata-cli`:

```python
from hepdata_cli.api import Client
client = Client()
query = 'ONNX'
inspire_ids = client.find(query, ids='inspire').split()
results = client.find(query, keyword='resources')
for i, result in enumerate(results):
    print(inspire_ids[i])
    for resource in result['resources']:
        if query in resource['description']:
            print(resource['description'])
            print(resource['type'])
            print(resource['url'].replace('landing_page', 'view'))
            print()
```

```
1869040
ONNX files for the neural networks for the EWK analysis
tgz
https://www.hepdata.net/record/resource/2677521?view=true

2043503
ONNX records of the NNs
gz
https://www.hepdata.net/record/resource/3170098?view=true
```

Note: this slide added *after* my talk (on 22nd June 2024) when I realised that the last item of the next slide is already satisfied by `hepdata-cli`.

OpenMAPP (03/2024 - 02/2026)

Tasks		
T1.1	**Extension of HEPData functionalities (M1–M6: responsible: 7; involved: 4,5,7)** - Add backend code to the HEPData system to store, retrieve, and be able to query data types beyond primary data-tables and uncertainties. - Build a HEPData mechanism for linking between data objects via internal DOIs, e.g. for associating theory predictions to measurement tables, or reference simulated event samples to their validation datasets. - Add a REST querying mechanism to HEPData, allowing interrogation of what resources of which types are associated with an analysis, and allowing each to be retrieved by a distinct URL. - Provide a programmatic interface to the REST API via the hepdata_cli Python package. The data types concerned by T1.1 include statistical models in JSON format, serialised machine-learning models (e.g., ONNX), binned and symbolic detector-response functions, combined sets of event-generator steering files for reference configurations and their corresponding reference outputs, etc. To be carried out by the UK partners, coordinated by partner 7 (Graeme Watt)	✅ Resources indexed ✅ Bidirectional linking ✅ Resources in JSON ✅ `hepdata-cli`

- Most of Task T1.1 now implemented, but open to feedback.

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).
- Infrastructure via CERN, development and support via Durham.
- Open development process via https://github.com/HEPData .
- New features since LHC Reinterpretation Forum in Dec 2022:
  - Numerous improvements to `hepdata_lib` helper library.
  - Default CC0 license with option to specify a different license.
  - Record-specific `<title>` tags to manage browser history.
  - Deferred loading in browser of YAML and text files larger than 1 MB.
  - Bidirectional linking possible between HEPData tables or records.
  - Resource metadata indexed and available in JSON format.
- OpenMAPP Task T1.1 implemented (?), but feedback welcome.