

Software Citation Tools, Technologies, and Best Practices

Matthew Feickert

(University of Wisconsin-Madison)

matthew.feickert@cern.ch

Software Citation and Recognition in HEP Workshop 2022

November 23rd, 2022



American Family Insurance
Data Science Institute
UNIVERSITY OF WISCONSIN-MADISON



Introduction and Overview

- In Tuesday's session, Daniel Katz already gave very nice high level overview of software citation **principles** and **tools**
- This is an **opinionated** summary of the tooling landscape and examples of workflows
 - Full disclosure: Opinions formed from pyhf development and from Scikit-HEP community discussions (c.f. [Eduardo's talk](#)).
- Meant to be recommendations to software developers on making your work as **easy to cite as possible**
 - These recommendations can transfer to experiment software as well



Daniel Katz's talk

Make clear how to cite in documentation

- The easiest, but least robust way: If you have a particular citation that you want people to use, put it **everywhere**
 - Version control repository README
 - Online software documentation (landing page, how to cite page)
 - Package distribution websites (e.g. PyPI)
- Having single source of truth for citations: version control repository that all other sources derive from.
- Make your citation preferences clear to the world and SEO. Do not rely on people emailing to ask (they shouldn't have to).



Use and Citations [Edit on GitHub](#)

Use and Citations

Citation

The preferred BibTeX entry for citation of [pyhf](#) includes both the [Zenodo](#) archive and the [JOSS](#) paper:

```
@software{pyhf,
  author = {Lukas Heinrich and Matthew Feickert and Giordon Stark},
  title = "{pyhf: v0.7.0}",
  version = {0.7.0},
  doi = {10.5281/zenodo.1169739},
  url = {https://doi.org/10.5281/zenodo.1169739},
  note = {https://github.com/scikit-hep/pyhf/releases/tag/v0.7.0}
}

@article{pyhf_joss,
  doi = {10.21105/joss.02823},
  url = {https://doi.org/10.21105/joss.02823},
  year = {2021},
  publisher = {The Open Journal},
  volume = {6},
  number = {58},
  pages = {2823},
  author = {Lukas Heinrich and Matthew Feickert and Giordon Stark and Kyle Cranmer},
  title = {pyhf: pure-Python implementation of HistFactory statistical models},
  journal = {Journal of Open Source Software}
}
```

[pyhf's "Use and Citations" page in documentation](#)

CITATION.cff

- Adopt the [Citation File Format](#) as a common standard and add a `CITATION.cff` to project repository
 - Human- and machine-readable file format in YAML
 - Has well defined, versioned schema
 - Convertible to other citation formats (BibTeX, CodeMeta, EndNote, RIS, schema.org, Zenodo, APA)

- Supported by [GitHub](#), [Zenodo](#), and [Zotero](#)!

- [Web tool initializer](#) for easily creating first `CITATION.cff`

- [Tooling for validation](#)

```
$ python -m pip install cffconvert
$ cffconvert --validate
Citation metadata are valid according to schema version 1.2.0.
```

```
cff-version: 1.2.0
message: "If you use this software, please cite it as below."
authors:
  - family-names: Druskat
    given-names: Stephan
    orcid: https://orcid.org/0000-0003-4925-7248
title: "My Research Software"
version: 2.0.4
doi: 10.5281/zenodo.1234
date-released: 2021-08-11
```

Example of minimal `CITATION.cff`

The screenshot shows a GitHub repository page for 'hainesr'. The commit history table is as follows:

Commit	Message	Time
db84466	Fix some minor issues with CFF fixtures.	11 days ago
	Remove the .ruby-* files from the repo.	last month
	Update README with new Model and File APIs.	
	Update the LICENCE and the file headers.	
	Turn on and fix rubocop Style/FrozenStringLiteralCom...	
	Add a code of conduct.	
	Update the CITATION.cff file to add a comment.	
	Update CHANGES.md and CITATION.cff for release.	
	Turn Coveralls reporting back on after move to Action...	
	Reference: new can now accept a block.	
	Reference: new can now accept a block.	
	Turn on and fix rubocop Style/FrozenStringLiteralCom...	
	Turn Coversalls reporting back on after move to Action...	
	Fix some minor issues with CFF fixtures.	
	Reference: new can now accept a block.	
	Turn on and fix rubocop Style/FrozenStringLiteralCom...	
	Turn Coversalls reporting back on after move to Action...	

The 'Cite this repository' modal window shows the following citation information:

Cite this repository
If you use this software in your work, please cite it using the following metadata. [Learn more](#)

APA **BibTeX**

Haines R. (2018). Ruby CFF Library (versio) [Software]. Zenodo. [DOI: 10.5281/zenodo.1234](#)

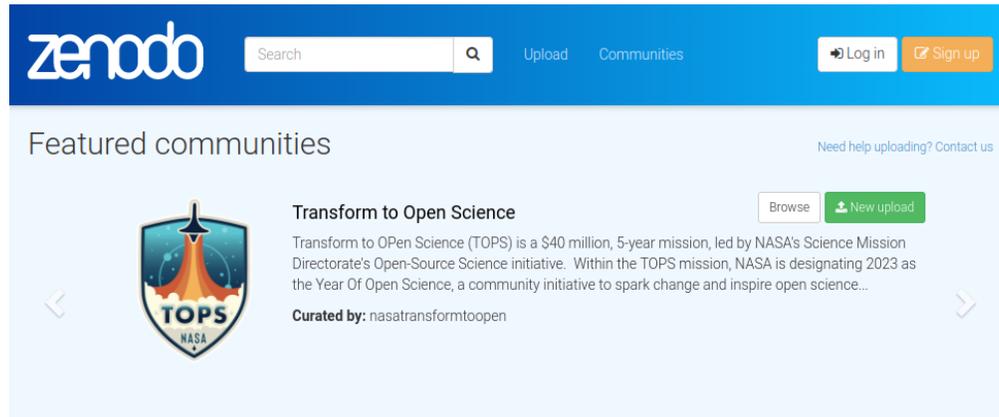
[View citation file](#)

CITATION.cff: How to keep up to date?

- As plain text, very easy to update version information when cutting a release
- Can use tool control of version update to make it easier
 - Example: `tbump`
 - `$ tbump <version target>`
- Also possible to have [automated version bump workflows](#) using continuous integration
- (Jumping ahead a slide) What about the Zenodo DOI?
 - For simplicity, use the project level DOI and not the version level DOI

```
cff-version: 1.2.0
message: "Please cite the following works when using this software."
type: software
...
title: "mylibrary: v1.2.3"
version: 1.2.3
doi: 10.5281/zenodo.1123456
repository-code: "https://github.com/myorg/mylibrary/releases/tag/v1.2.3"
url: "https://mylibrary.readthedocs.io/en/v1.2.3/"
```

Zenodo



zenodo Search Upload Communities Log in Sign up

Featured communities [Need help uploading? Contact us](#)

 **Transform to Open Science** [Browse](#) [New upload](#)

Transform to Open Science (TOPS) is a \$40 million, 5-year mission, led by NASA's Science Mission Directorate's Open-Source Science initiative. Within the TOPS mission, NASA is designating 2023 as the Year Of Open Science, a community initiative to spark change and inspire open science...

Curated by: [nasatransformtoopen](#)

Recent uploads

November 20, 2022 (v141) Dataset Open Access [View](#)

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

 Banda, Juan M.;  Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena;  Chowell, Gerardo

Version 141 of the dataset. MAJOR CHANGE NOTE: The dataset files: full_dataset.tsv.gz and full_dataset_clean.tsv.gz have been split in 1 GB parts using the Linux utility called Split. So make sure to join the parts before unzipping. We had to make this change as we had huge issues uploading...

Uploaded on November 21, 2022
141 more version(s) exist for this record

Need help?

[Contact us](#)

Zenodo prioritizes all requested related to the COVID-19 outbreak.

We can help with:

- Uploading your research data, software, preprints, etc.
- One-on-one with Zenodo supporters.
- Quota increases beyond our default policy.
- Scripts for automated uploading of larger datasets.

- Open source (but your files can be closed access)
- Versioned archival of everything: code, documents, data products, data sets

Why use Zenodo?

- **Safe** — your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
- **Trusted** — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- **Citeable** — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.
- **No waiting time** — Uploads are made available online as soon as you hit publish, and your DOI is registered within seconds.
- **Open or closed** — Share e.g. anonymized clinical trial data with only medical professionals via our restricted access mode.
- **Versioning** — Easily update your dataset with our versioning feature.
- **GitHub integration** — Easily preserve your GitHub repository in Zenodo.
- **Usage statistics** — All uploads display standards compliant usage statistics

Zenodo: DOI minting made easy

- Everything on Zenodo has a DOI
 - Provides both a **project** DOI (resolves to latest) and **version specific** DOI
- Enable it to [automatically preserve work from GitHub](#) (can also directly upload, but lose out on automation)
 - Benefit from having a DOI for **every version** regardless of software paper landscape state
- Once you have a DOI, put it **everywhere** (again)
 - Recommend sharing the project DOI and letting users select a specific version if they want it

 **GitHub Repositories** (updated 51 minutes ago)  Sync now ...

Get started

- ### 1 Flip the switch

Select the repository you want to preserve, and toggle the switch below to turn on automatic preservation of your software.

ON
- ### 2 Create a release

Go to GitHub and [create a release](#). Zenodo will automatically download a .zip-ball of each new release and register a DOI.
- ### 3 Get the badge

After your first release, a DOI badge that you can include in GitHub README will appear next to your repository below.

DOI `10.5281/zenodo.8475`
(example)

Zenodo + CITATION.cff

CITATION.cff used by Zenodo importer to fully define Zenodo archive metadata

```
71 lines (71 loc) · 2.47 KB
1  cff-version: 1.2.0
2  message: "Please cite the following works when using this software."
3  type: software
4  authors:
5  - family-names: "Heinrich"
6    given-names: "Lukas"
7    orcid: "https://orcid.org/0000-0002-4048-7584"
8    affiliation: "Technical University of Munich"
9  - family-names: "Feickert"
10   given-names: "Matthew"
11   orcid: "https://orcid.org/0000-0003-4124-7862"
12   affiliation: "University of Wisconsin-Madison"
13  - family-names: "Stark"
14   given-names: "Giordon"
15   orcid: "https://orcid.org/0000-0001-6616-3433"
16   affiliation: "SCIPP, University of California, Santa Cruz"
17  title: "pyhf: v0.7.0"
18  version: 0.7.0
19  doi: 10.5281/zenodo.1169739
20  repository-code: "https://github.com/scikit-hep/pyhf/releases/tag/v0.7.0"
21  url: "https://pyhf.readthedocs.io/en/v0.7.0/"
22  keywords:
23  - python
24  - physics
25  - statistics
26  - fitting
27  - scipy
28  - numpy
29  - tensorflow
30  - pytorch
31  - jax
32  - auto-differentiation
33  license: "Apache-2.0"
```

September 24, 2022 Software Open Access

scikit-hep/pyhf: v0.7.0

Lukas Heinrich, Matthew Feickert, Giordon Stark
pure-Python HistFactory implementation with tensors and autodiff

10,950 views 86 downloads
[See more details...](#)

Available in

Indexed in

Publication date: September 24, 2022
DOI: [10.5281/zenodo.7110486](https://doi.org/10.5281/zenodo.7110486)
Keyword(s): [physics](#) [statistics](#) [fitting](#) [scipy](#) [numpy](#) [tensorflow](#) [pytorch](#) [jax](#) [auto-differentiation](#)
Related identifiers: Supplement to <https://github.com/scikit-hep/pyhf/tree/v0.7.0>
License (for files): [Apache License 2.0](#)

Versions
Version v0.7.0 Sep 24, 2022

Zenodo: Communities allow archival collections

 Communities created and curated by Zenodo users

PyHEP

Showing 0 to 5 out of 5 communities.

Sort by ▾

PyHEP 2018 Workshop

View

"Python in HEP" 2018 Workshop held on July 7-8 2018 in Sofia, Bulgaria.
Agenda: <https://indico.cern.ch/event/694818/>.

Curated by: eduardo-rodrigues

PyHEP 2021 Workshop

View

PyHEP 2021 Workshop ("Python in HEP") held on July 5-9 2021 as a virtual event. Agenda: <https://indico.cern.ch/e/PyHEP2021>.

Curated by: eduardo-rodrigues

PyHEP 2022 Workshop

View

PyHEP 2022 Workshop ("Python in HEP") held on September 12-16 2022 as an online event. Agenda: <https://indico.cern.ch/e/PyHEP2022>.

Curated by: eduardo-rodrigues

PyHEP 2020 Workshop

View

"Python in HEP" 2020 Workshop held on July 13-17 2020 as a virtual event. Agenda: <https://indico.cern.ch/e/PyHEP2020>.

Curated by: eduardo-rodrigues

PyHEP 2019 Workshop

View

"Python in HEP" 2019 Workshop held on October 16-18 2019 at The Cosener's House, Abingdon, U.K.
Agenda: <https://indico.cern.ch/e/PyHEP2019>.

Curated by: eduardo-rodrigues

The ATLAS Experiment at CERN

Recent uploads

Search The ATLAS Experiment at CERN



View

March 16, 2022 (1.1.0) Software Open Access

SimpleAnalysis

Atlas Collaboration;

SimpleAnalysis (SA) is an analysis framework in C++ designed to run on the output of event generators (generator-level information, or truth) which is used in most ATLAS Supersymmetry (SUSY) results. Similar in scope to Rivet, SA is a software framework running on data formats belonging to the ATLAS

Uploaded on March 17, 2022

1 more version(s) exist for this record

March 11, 2022 (v4.10/sw4.2) Software Open Access

View

The ATLAS FELIX Project

ATLAS TDAQ Collaboration;

FELIX is a key component of the new readout architecture for the ATLAS experiment at CERN. It consists of commodity servers hosting PCIe cards, which receive data from detector and trigger electronics over optical links and transfer them to the host system via Direct Memory Access (DMA). From there,

Uploaded on March 11, 2022

October 5, 2021 (1.0) Software Open Access

View

FastCaloGAN Training Project

ATLAS collaboration;

FastCaloGAN is the tool developed and used by the ATLAS experiment at CERN for simulating particle showers in the calorimeter system. FastCaloGAN exploits Generative Adversarial Networks that are trained on samples that will be provided in ATLAS open data website, <http://opendata.cern.ch>. A detaille

Uploaded on October 21, 2021

September 1, 2021 (1.11.2) Software Open Access

View

BootstrapGenerator

New upload

Community



The ATLAS Experiment at CERN

ATLAS is one of the four major experiments at the Large Hadron Collider (LHC) at CERN. It is a general-purpose particle physics experiment run by an international collaboration and, together with CMS, is designed to exploit the full discovery potential and the huge range of physics opportunities that the LHC provides.

ATLAS' scientific exploration uses precision measurement to push the frontiers of knowledge by seeking answers to fundamental questions such as: What are the basic building blocks of matter? What are the fundamental forces of nature? Could there be a greater underlying symmetry to our universe?

See here for more: <https://atlas.cern>

Curated by:
atlasexperiment

Curation policy:
Only official software, approved by the

Proving a citation information from APIs

- In addition to providing standard formats, providing users a language API or CLI API to get the citation information for the version of the tool is helpful
 - User doesn't have to check if the information they find online matches their version.
- Historically, this was done by printing a banner with citation or copyright information when the library is used
 - This should **not** be done now. This creates noise for users and if multiple tools did this your terminal would get filled.
 - Most libraries that used to do this have now abandoned this approach.
- Opinion: There are tools in broader scientific ecosystem that provide citation information for their dependencies as well. While very conscientious, I think this is **unnecessary** and can be confusing to users.

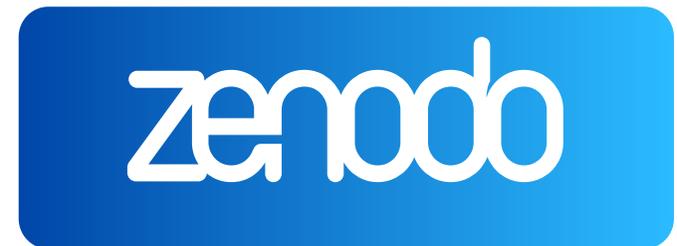
```
# CLI API
$ mytool --citation
$ mytool --cite
```

```
# Python API
import mytool
mytool.utils.citation()
```

Example APIs

Summary

- Build community practices on top of **established standards**
- If citation of your software is important to you, **make it easy** for a user to find your citation information
- Modern standards like `CITATION.cff` allow for **single source of citation information** that can be exported as needed
- Long term archives + **FAIR practices**
 - Zenodo provides automatically release information each release



Backup

Does any of this actually work?

As mentioned, these opinions have been formed from developing pyhf, and the citation count for the [JOSS paper](#) has increased each year.

The screenshot shows the INSPIRE HEP website interface. At the top, there is a navigation bar with the INSPIRE HEP logo, a search bar containing the word 'literature', and links for 'Help' and 'Submit'. Below the navigation bar, there are tabs for 'Literature', 'Authors', 'Jobs', 'Seminars', 'Conferences', and 'More...'. The main content area displays the title 'pyhf: pure-Python implementation of HistFactory statistical models' by Lukas Heinrich (CERN), Matthew Feickert (Illinois U., Urbana), Giordon Stark (UC, Santa Cruz, Inst. Part. Phys.), and Kyle Cranmer (New York U.), published on Feb 4, 2021. It also lists '2 pages', 'Published in: J. Open Source Softw. 6 (2021) 58, 2823', 'Published: Feb 4, 2021', 'DOI: 10.21105/joss.02823', and 'View in: CERN Document Server'. At the bottom of the article, there are icons for 'pdf', 'cite', and 'claim', along with a 'reference search' icon and '36 citations'. To the right of the article is a line graph titled 'Citations per year' showing an upward trend from 1 citation in 2020 to 21 citations in 2022.

Year	Citations
2020	1
2021	14
2022	21

