Bidirectional Paper-Repository Tracing in Software Engineering

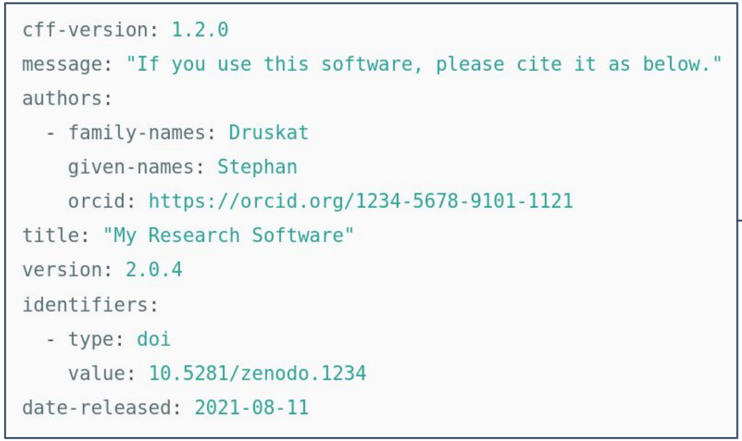
Daniel Garijo[1] , Esteban Gonzalez[1] , Miguel Arroyo[1] , Christoph Treude[2], Nicola Tarocco[3]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid
[2] The University of Melbourne
[3] CERN

✉ daniel.garijo@upm.es
🐦 @dgarijov

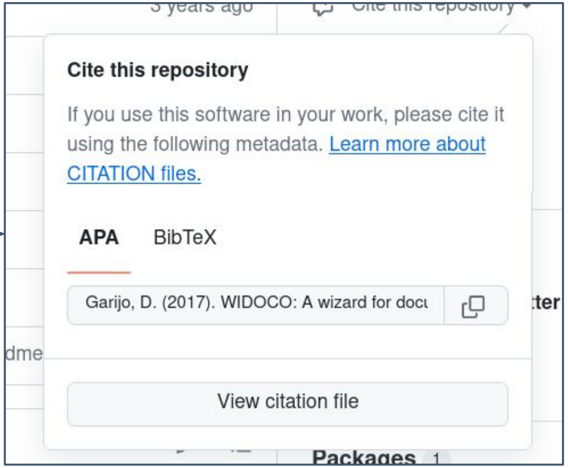
Software citation is on the rise!

citation file format



```
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/1234-5678-9101-1121
title: "My Research Software"
version: 2.0.4
identifiers:
  - type: doi
    value: 10.5281/zenodo.1234
date-released: 2021-08-11
```

GitHub

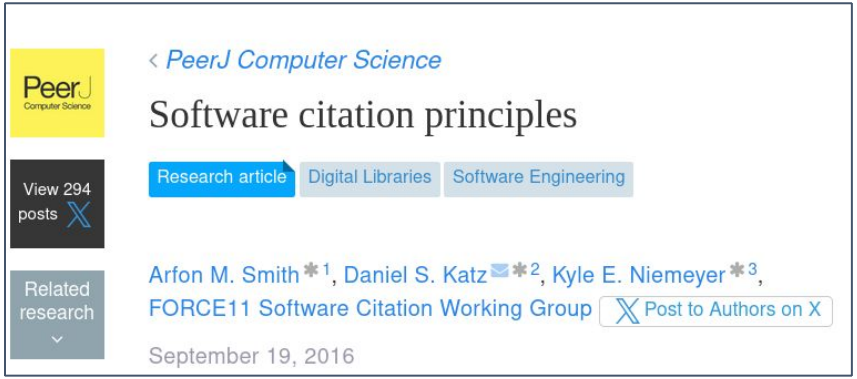


Cite this repository

If you use this software in your work, please cite it using the following metadata. Learn more about CITATION files.

Garijo, D. (2017). WIDOCO: A wizard for docu

Software citation principles



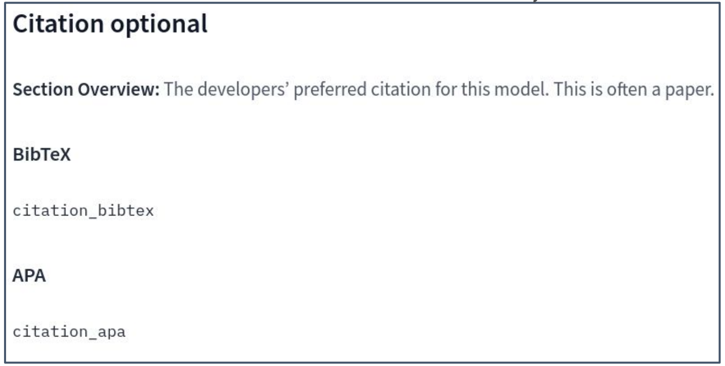
< *PeerJ Computer Science*

Software citation principles

Arfon M. Smith [1], Daniel S. Katz [2], Kyle E. Niemeyer [3], FORCE11 Software Citation Working Group

September 19, 2016

Others: Model Cards, PwC



Citation optional

Section Overview: The developers' preferred citation for this model. This is often a paper.

BibTeX

`citation_bibtex`

APA

`citation_apa`

Our goal

- Find **the** software implementation(s) associated with a paper

- Find how authors adopt citation practices in their tool implementations

- Assess current software metadata practices (FAIR principles)

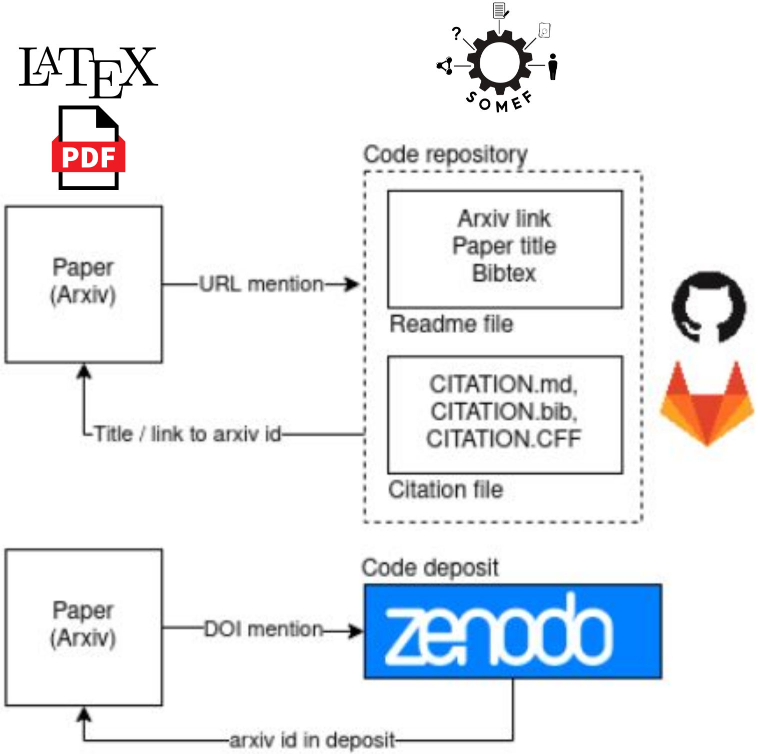
Challenge: Diverse citation practices

- **Main way in which software is cited**:
  - Footnotes
  - Inline URLs
  - Citations (as if it was a paper)

- **Cited artefacts**:
  - Zenodo DOIs
  - GitHub repos
  - Project sites

- **Additional challenges**:
  - Links in the pdf (which can be clicked) but no citation reference
  - It may be impossible to discern from the context that the cited resource is the paper implementation
  - Links to data, experiment evaluations and software code are provided **as code**
  - Many, many tools that are **reused** in a paper (but not proposed) are cited
  - Arxiv papers may link to implementations that are accepted in **conference papers**

Bi-directional paper-repository traceability methodology



LATEX

PDF

SOMEF

Code repository

Paper (Arxiv)

URL mention

Arxiv link
Paper title
Bibtex

Readme file

Title / link to arxiv id

CITATION.md,
CITATION.bib,
CITATION.CFF

Citation file

Code deposit

Paper (Arxiv)

DOI mention

zenodo

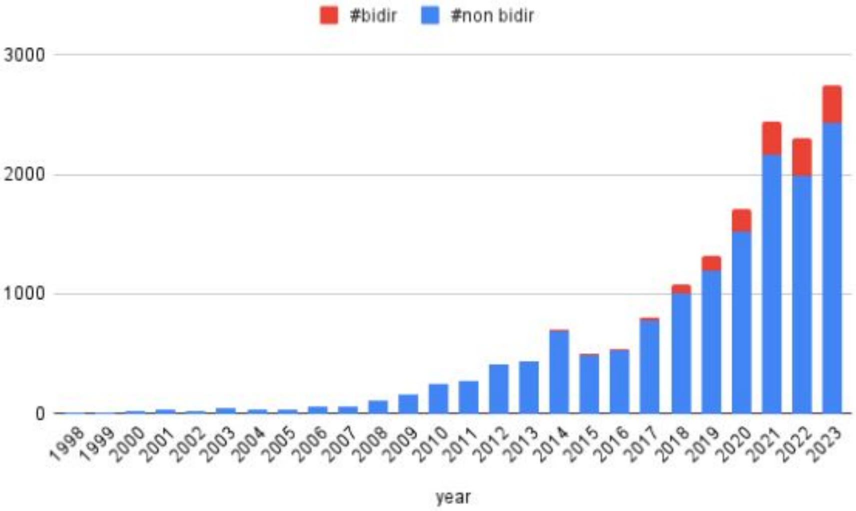
arxiv id in deposit

Our method:

- Precise, robust method for **finding bidirectional paper-links**
  - Caveat: may miss unidirectional
- Explores Zenodo links that have a repository and link back to the paper
- Similarity based on DOI, mention references of paper title found in README.
- Validation over 150 papers:
  - PDF: P: 1, R: 0.94, **F1: 0.97**
  - Latex: P:1, R: 0.7, **F1: 0.83**

https://github.com/SoftwareUnderstanding/RSEF (PDF)
https://github.com/ctreude/SoftwareImpactHackathon2023_BiDirectional (Latex)

Results



14760 PDFs and 10826 Latex files
in **Cs.Se category** (Software Engineering)

Pipeline	**GitLab**	**Zenodo**	**GitHub**	**Total**
PDF	7	273	1159	1439
Latex	0	40	638	678

Citation practices	Description	Bibtex	CFF	**Title**
Bidirectional	759	307	49	353

Software citation is becoming **increasingly important**, but best practices are not yet followed

Conclusions and Future work

- Extracting software implementations from papers accurately **may be challenging**
  - Diverse cited artifacts
  - Diverse citation practices
  - Technical challenges

- Robust pipeline for mining software implementations

- Future work will address
  - **Unidirectional** paper-repositories
  - Other domains

- Ack:

Comunidad
de Madrid
Dirección General de Investigación
e Innovación Tecnológica
CONSEJERÍA DE CIENCIA,
UNIVERSIDADES E INNOVACIÓN

Data

Corpus

Results

Code

PDF

Latex

How are you asking others to cite your software?

How do you cite other software in your work?

Daniel Garijo[1] **,** Esteban Gonzalez[1] , Miguel Arroyo[1] , Christoph Treude[2], Nicola Tarocco[3]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid
[2] The University of Melbourne
[3] CERN

✉ daniel.garijo@upm.es
🐦 @dgarijov