Annotation guidelines

Stephan Druskat[1,2], Neil P. Chue Hong[3], Sammie Buzzard[4], and Olexandr Konovalov[5]

[1]German Aerospace Center (DLR), Institute for Software Technology, Berlin, Germany
[2]Humboldt-Universität zu Berlin, Department of Computer Science, Berlin, Germany
[3]EPCC, University of Edinburgh, Edinburgh, United Kingdom
[4]School of Earth and Environmental Sciences, Cardiff University, Cardiff, United Kingdom
[5]School of Computer Science, University of St Andrews, St Andrews, United Kingdom

Corresponding author:
Stephan Druskat[1]

Email address: stephan.druskat@dlr.de

ABSTRACT

Annotation guidelines for the sample data described in `Stephan Druskat, Neil P. Chue Hong, Samiie Buzzard, Olexandr Konovalov, Patrick Kornek. Don't mention it: challenges to using software mentions to investigate citation and discoverability.`

ANNOTATION GUIDELINES

For each software mention in the sample,

1. Resolve the first identifier for the publication in a web browser.

   1.1. If the publication is a preprint, use the next identifier if available.

   1.2. If the only available identifier is for a preprint, use the preprint.

2. Open the PDF for the publication.

   2.1. If you cannot access the PDF due to a paywall, use the next identifier.

   2.2. If there is no next identifier, use Unpaywall[1] to access an open version of the publication, or ask a co-author to retrieve the publication.

3. Search for the exact mention string in the PDF.
4. Verify for each search result that it is the exact search string. Note that:

   4.1. The mention string may be a substring of the complete software name (due to line breaks, composite names, etc.).

   4.2. There may be multiple software packages mentioned with similar names.

5. Annotate the quality of the mention retrieval according to Table 1.
6. Identify the best mention and annotate the mention type.

   6.1. Identify the best mention by adherence to the software citation principles.

[1] https://unpaywall.org/

Code	Name
Y	Yes, name was correctly and completely retrieved from the publication for the dataset.
N	No, name was NOT correctly and completely retrieved from the publication for the dataset.

Table 1. Annotations for quality of the mention extraction/retrieval.

6.2. The *Order* column in Table 2 encodes the quality of the mention (from 1 = best to 6 = worst) by principles:

- Importance is always the best. Citation of project name or website is better than citation of a publication. (Importance, Accessibility)
- Citation of a publication is better than citation of a user manual. (Credit)
- URLs in text are second best. (Accessibility)
- Instrument-like citation is better than name-only mention. (Accessibility)
- Name-only mentions are better than mention without name.

6.3. Only use mentions matching the exact mention string, including capitalization.

6.4. Only URLs found in the same paragraph as the mention, or in a footnote that is called from the same paragraph, shall be annotated with `URL`.

6.5. Citations to references must appear within the boundaries of the sentence that includes the mention.

6.5.1. Examples for citations to process:

- "We used SOFTWARE [1] for the analysis."
- "We used SOFTWARE for the analysis [1]."
- "We used SOFTWARE for the analysis. [1]"

6.5.2. Example for citations to ignore:

- "We used SOFTWARE and Otherthing for the analysis. We refuted the null hypothesis. The data provided evidence for something [1, 2]."

7. Annotate the quality of the mention (Table 3).

7.1. Differentiate between mention types `NA` and `SN`.

7.1.1. If it is clear that the authors considered the mentioned entity software, annotate as `SN`. Examples: listed as "computational method", compared with other software.

7.1.2. If still unclear, discuss with other annotators.

7.1.3. If still unclear, annotate as `UN`.

8. Annotate other layers.

REFERENCES

Howison, J., & Bullard, J. (2015). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, *67*(9), 2137–2155. https://doi.org/10.1002/asi.23538

Code	Name	Definition	Order
PUB	Cite to publication	Cites a paper/monograph primarily describing the mentioned software (NOT a review paper comparing different software), as it would for non-software cites. For non-software mentions, we don't judge the suitability of the referenced work.	2
PRO	Cite to project name or website	Cites the project name or website via a "fake" reference.	1
URL	URL in text	URL in text or in footnote	4
MAN	Cite to user manual		3
INS	Instrument-like	Mention software in a manner similar to scientific instruments or materials, typically mentioning the name in text followed by the author or company and a location in parentheses.	5
NAM	In-text name mention only		6
NOT	Not even name mentioned		7

Table 2. Annotations for mention types following Howison and Bullard (2015).

Code	Name
SC	Software where a direct link to a code repository or distribution repository landing page (e.g., CRAN, PyPI) can be found in the mentioning paper, and the page includes author/version/ license metadata.
SP	Software where a link to another website can be found in the mentioning paper and that website provides access to the source code, but the website does not provide author/version/license metadata.
SN	Software but no link to a code repository or website providing access to the source code can be found in the mentioning paper. Annotate as SN even if the reference is to a software paper that does include a link to a source code repository.
NA	Not software (only annotate this, retrieval quality and confidence)
UN	Other classification - unknown/needs further investigation, e.g., unclear from the information in the paper whether this is software or not.

Table 3. Annotations for quality of the mention itself.