Don’t mention it v2 - Annotation Guidelines

Version 2022-01-24

Annotation workflow

The sample is annotated on 2 (or 3, if we decide to do QA again) layers:

1. Citation type
2. Accessibility and principledness

Both layers partially follow the same annotation workflow, but differ in classification.

Workflow

1. Open the <sample CSV> file in an editor of your choice (LibreOffice Calc, Excel, …).
2. For each row of the spreadsheet
    - 2.1. **Open a URL from the rand_url field in a browser** (prefix DOIs with https://doi.org/), following these rules:
        - 2.1.1. If the first URL in the field is to a **preprint** (e.g., on the ArXiv), use the leftmost one that is **not** to a preprint.
        - 2.1.2. If the first URL in the field is **not** to a preprint, use it.
        - 2.1.3. Use a preprint URL if it is the only URL in the field.
    - 2.2. Look for a link to a PDF file and **open the PDF**
        - 2.2.1. If you cannot access the PDF due to it being paywalled, try the next link in the rand_url field, etc.
        - 2.2.2. If you cannot access a PDF from any of the URLs in the rand_url field, use Unpaywall to search for a freely available version/preprint. Note this in the comments field.
    - 2.3. **Search** for the exact mention string (copy and paste from the spreadsheet perhaps)
    - 2.4. **Verify** for each search result that it is actually == the search string
        - Sometimes the mention string may be a substring of the complete software name (due to line breaks, composite names, etc.), sometimes there are more than one software packages mentioned with similar names, make sure you pick only the correct ones
        - Note wrongly represented software names in the spreadsheet
    - 2.5. **Pick the best** mention/citation
        - For the definition of “best”, see section below
        - When searching for a reference, the boundaries of the sentence in which the mention appears are the limits for taking references into account (i.e., classify as PRO or PUB):
            - Counted as PUB/PRO:
                - “We used SOFTWARE [1] for the analysis.”
                - “We used SOFTWARE for the analysis [1].”

        - “We used SOFTWARE for the analysis. [1]”
    - NOT counted as PUB/PRO:
        - “We used SOFTWARE and Otherthing for the analysis. In the process, we found that our initial assumptions were correct. The processed dataset provided clear evidence for something [1, 2].”
    - (Note that the best citation may not be the one that is included in the CORDS-19 SMD, we can’t verify that, and will explain that in the paper)

2.6. **Categorize** it using the tagset table

2.7. Lather, rinse, repeat

What is the “best” citation?

- Categorization by quality depends on adherence to the software citation principles/fulfillment of the functions of citation
- Quality is encoded in the **Order** column in the tagset table for the main layer (ordered from 1 = best to 6 = worst)
- Importance trumps everything else, i.e., a reference item is always the best
    - A cite to project name or website (PRO) trumps a cite to a paper (PUB), because of the Importance principle ("cite the software itself") and because it may allow better accessibility (if a URL is provided)
    - A cite to a publication (PUB) trumps a cite to a user manual (MAN), because of better credit
- URL in text (URL) is second, because it enables accessibility (perhaps even of the source code)
- Instrument-like may allow accessibility, but usually not of the source code, but it's better than ...
- In-text name mention only, which is still better than ...
- nothing

Annotation tagset

- We use the categories from Howison & Bullard 2015 (PDF)
- There are 7 annotation values on the main layer (from Table 6., p. 7)

Layer tagset tables on the following pages

MAIN layer

Code	Name	Definition	Examples from our corpus	Order
PUB	Cite to publication	Cites a paper/monograph **primarily describing the mentioned software**, as it would for non-software cites	*the full path of extracting the RoI from the FPN layer using RoIAlign ([HGDG17]), predicting bounding box coordinates,* + [HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. InProceedings of the IEEEinternational conference on computer vision, pages 2961–2969, 2017.	2
PRO	Cite to project name or website	Cites the project name or website via a "fake" reference	*TekStack [7]* + [7] "TEKStack Health - COVID-19 Research Portal." [Online]. Available:https://covid-research.tekstackhealth.com/	1
URL	URL in text	URL in text or in footnote	*the SQID[29] KG* + [29]https://tools.wmflabs.org/sqid/	4
MAN	Cite to users manual		From H & B: *. . . as analyzed by the BIAevaluation software (Biacore, 1997).* + Biacore, I. (1997). BIAevaluation Software Handbook,version 3.0 (Uppsala, Sweden: Biacore, Inc)	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	*Data were entered in two Microsoft Access databases (Microsoft Corp, Va., USA) for patient's*	5
NAM	In-text name mention only		*and the PMN program of the Spectrasoftware version 4.7.*	6
NOT	Not even name mentioned			7

Note: When a publication is cited, but the publication is not a software paper or a paper primarily describing the mention ed software, this should **NOT** be classified as PUB. Instead, classify with the most suitable lesser code, e.g., NAM.

QA layer

Code	Name
SO	Software where a link to a code repository can be found
SC	Software but no link to a code repository can be found
ST	Typo but the mention is to software
SF	Specific function / subroutine in a larger software package or library with a different name
NA	Not software but correctly spelt
NT	Not software but incorrectly spelt
UN	Other classification - unknown / needs further investigation

Mention retrieval QA layer (**QA_retrieval**)

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

Examples for N:

- CSM includes “Snap” for “Snap Chat”
- CSM includes “SciKit” and “Learn” for “SciKit Learn”
- CSM includes “Hyphenated-” for “Hyphenated-Softwarename”

Preprint layer (**preprint**)

Code	Name
Y	Yes, publication containing the software mention is a preprint.
N	No, publication containing the software mention is NOT a preprint.

Software paper layer (**software_paper**)

Code	Name
Y	Yes, publication containing the software mention is a paper primarily describing the mentioned software / a software paper.
N	No, publication containing the software mention is NOT a paper describing primarily the mentioned software / a software paper.

Confidence layer (**confidence**)

Code	Name
Y	Yes, the annotator is confident that their annotations are correct.
N	No, the annotator is NOT confident that their annotations are correct.

Further columns

- **annotator**: identifies the annotator(s) of this item (SD, NCH, SB, AK, PK)
- **comments**: Free text comments

Layers for single analyses

DETAILS layer (Stephan)

Code	Name	Examples from our corpus	Notes
VER	Version information in reference	Shrikumar A, Tian K, AvsecŽ, Shcherbina A, Banerjee A, Sharmin M, et al.Technical note on transcription factor motif discovery from importancescores (TF-MoDISco) version 0.5.1.1; 2018. arXiv preprint , arXiv:1810.04805.	E.g., Software 4.11.1, etc.
REV	Version information near mention	*and the PMN program of the Spectrasoftware version 4.7.*	E.g., Software 4.11.1, etc.
CRE	Creditable author information in reference	Shrikumar A, Tian K, AvsecŽ, Shcherbina A, Banerjee A, Sharmin M, et al.Technical note on transcription factor motif discovery from importancescores (TF-MoDISco) version 0.5.1.1; 2018. arXiv preprint , arXiv:1810.04805.	
ERC	Creditable author information near mention		
REP	Link to community repository version in reference	Almeida, A.; Loy, A.; Hofmann, H. Qqplotr: Quantile-Quantile Plot Extensions for “ggplot2”. 2020. Available online: https://cran.r-project.org/package=qqplotr (accessed on 11 June 2020).	
PER	Link to community repository version near mention		
LIN	Link to source code in reference		
NIL	Link to source code near mention	freely available Kontaminant tool developed at The Genome Analysis Centre (TGAC), http://www.tgac.ac.uk/kontaminant/ or https://github.com/TGAC/kontaminant [42].	
SOF	Reference is to software itself	TEKStack Health - COVID-19 Research Portal.” [Online]. Available:https://covid-research.tekstackhealth.com 21.Sergeant, ESG. Epitools Epidemiological Calculators. Ausvet. Available at: http://epitools.ausvet.com.au.; 2018.	
UNI	Unique software (version) identifier in reference	E.g., a DOI to a version of the software, a Handle/RRID, an SWH ID	
INU	Unique software (version) identifier near mention		
PSV	Persistent software version link in reference		
VSP	Persistent software version link near mention		
0	No information		