Software Citation in the Digital Humanities

Jettka, Daniel

daniel.jettka@uni-paderborn.de
University of Paderborn, Germany

Henny-Krahmer, Ulrike

ulrike.henny-krahmer@uni-rostock.de
University of Rostock, Germany

Ferger, Anne

anne.ferger@uni-paderborn.de
University of Paderborn, Germany

Alvares Freire, Fernanda

fernanda.freire@uni-rostock.de
University of Rostock, Germany

Introduction

In contemporary humanities research and DH particularly, the use, adaption, and development of software plays a central role in acquiring, enriching, analyzing, and publishing research data. Published software is, in principle, citable, like published datasets or classic text publications, and should be referenced appropriately for good scientific practice to make software discoverable, make research results transparent, and give credit to software developers (Smith et al. 2016, Lamprecht et al. 2020).

This contribution reviews the practice of software citation in the DH. [1] The analysis is based on DH conference abstracts and anannotation model developed from existing software citation recommendations (for preliminary work see Henny-Krahmer/Jettka 2021, 2022). Through semi-automatic annotation and analysis, software citations are examined in the abstracts (and the information given or omitted with the citations).

Recommendations for software citation

Recommendations for software citation already exist, including Jackson (n.d.), Smith et al. (2016), Chue Hong et al. (2019a, 2019b), Druskat (2021a, 2021b) on software citation in particular, and Anzt et al. (2021) and Jettka/Henny-Krahmer (2022), amongst others, as general guides for sustainable software development. They suggest what references to software in scholarly texts should contain, especially in the form of bibliographic citations. Recommendations for software citations also come from general citation styles (including MLA or APA) or from developers themselves. In line with the existing software citation principles, we formulate the following criteria to characterize mentions of research software in DH abstracts, modeled as a TEI taxonomy and available on GitHub (see Jettka et al. 2022 and Henny-Krahmer/Jettka 2021 for details):

- **Bib.Soft:** bibliography entry for the software itself
- **Bib.Ref:** bibliography entry for reference publication about the software, e.g. journal article, book, manual
- **Name.Only:** software name only
- **Agent**: developers/responsible persons are named
- **URL**: citation contains a URL to the software itself
- **PID**: citation contains a persistent identifier for the software itself
- **Ver**: citation indicates a specific software version

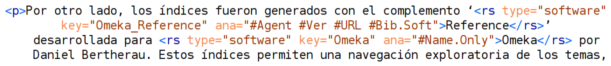
Data and Methods

Corpus

The corpus analyzed comprises volumes of DH conference abstracts from 2015 –2020, which are accessible on GitHub or the ADHO conference pages in TEI format. [2]There are 1,887 abstracts for panels, workshops, posters, and talks (approx. 2.7 million tokens and 100k types), of which we randomly selected 156 abstracts for manual annotation, making sure all languages (English, Spanish, French, German, Italian, Portuguese) and all relevant years are included.

Annotation of software citations

We opted for semi-automatic annotation to examine the citation modes as closely as possible. [3] Of the selected abstracts, 87 contained software citations, and 69 contained none. When annotating software citations, we broadly define software as any type of computer program, including desktop and web applications, server software, plugins, and extensions and sets of scripts. [4] Using the TEI taxonomy, software citations were annotated in the TEI files of the DH abstracts (see Fig. 1), and the cited software was collected in a central index of software names.



```
<p>Por otro lado, los índices fueron generados con el complemento ‘<rs type="software"
        key="Omeka_Reference" ana="#Agent #Ver #URL #Bib.Soft">Reference</rs>’
    desarrollada para <rs type="software" key="Omeka" ana="#Name.Only">Omeka</rs> por
    Daniel Bertherau. Estos índices permiten una navegación exploratoria de los temas,
```

Fig. 1: Example of annotated software citation

Citation types (except for Name.Only) are not exclusive, and multiple citation types can be associated with one specific software. For example, a software may have been cited in a paper with both Bib.Soft and Bib.Ref. The software list, TEI taxonomy, annotated TEI documents, the RELAX NG schema that the annotated documents can be validated with, and the analysis data are available in Jettka et al. (2022).

Results and Discussion

Software is often cited multiple times in an abstract but understandably not fully cited at each of its mentions. Thus the presence of citation components for each software is counted once per paper rather than per citation (e.g., if a URL is used as a software citation once in the paper, this criterion is counted as met). The distribution (see Fig. 2) is based on a total of n=266, representing the single-counted citations of a particular software in a particular paper.

Citation type	Abs. frequency (n=266)	Rel. frequency (in %)
Bib.Soft	16	6.02
Bib.Ref	62	23.31
Name.Only	187	70.3
Agent	11	4.14
URL	52	19.55
PID	1	0.38
Ver	9	3.38

Fig. 2: Frequency of software citations

The results show that in 187 (70.3%) cases where software is mentioned, none of the considered citation methods is used. The most frequent method is citing a reference publication about the software (23.31%), often providing a URL to the software or a website about the software (19.55%). Direct, long-term citation via bibliographic records for the software or using a persistent identifier is relatively rare. Software versions are also rarely given, and the people who developed the software are not properly credited. There seems to be only little to no change in the awareness of the need for sustainable citation of research software and the citation practice over the years (see Fig. 3). Significant statements, however, require a larger-scale study.

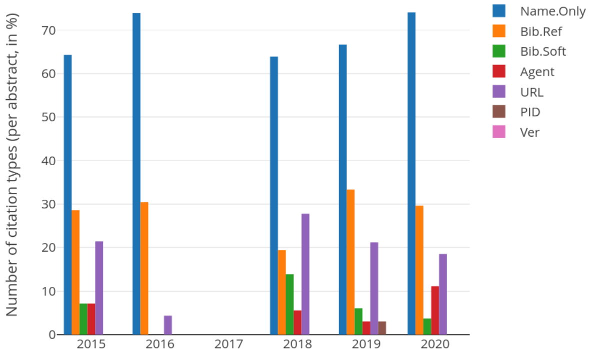


Fig. 3: Relative frequency of software citations per year

Conclusions

This paper presented criteria for analyzing software citations in the DH abstracts of 2015 –2020 which were formulated and represented in the form of a TEI taxonomy. Based on the taxonomy, semi-automatic annotation of the software and their citations in the DH abstracts was carried out.

The present software citation practice in the DH clearly points to a need for improvement, both with regard to the use of bibliographic entries for research software itself and the use of persistent identifiers, as well as with regard to the naming of responsible persons and institutions, whose achievements should be recognized accordingly. There is still a need for action both on the part of citation practice, i.e., users, and on the part of citation recommendations for software projects, i.e., developers and operators, and publishers who provide citation guidelines.

Notes

1. See Howison/Bullard (2016) for a similar study on software citation in biological research publications.
2. The 2015, 2016, 2018 abstracts were originally published on ADHO's GitHub page (https://github.com/ADHO/, licensed under CC BY), 2019 and 2020 on the respective conference pages. They were, however, retrieved in TEI format from the ToolXtractor's GitHub page: https://github.com/lehkost/ToolXtractor (licensed under Apache 2.0). We did not include 2017 abstracts, as they were not directly available in TEI format.
3. Complementary to manual annotation, we used ToolXtractor (Barbot et al. 2019; 2021), based on a manually created software list, to find potential software citations that were only considered after manual post-processing. Additionally to manual approaches, there are fully automatic ones to detect software mentions (see, for instance, Krüger/Schindler (2020), Zarei et al. (2022)), which we did not use.
4. For more specific definitions of research software see Hettrick et al. (2014) and Homburg et al. (2020).

Bibliography

Anzt, Hartwig / Bach, Felix / Druskat, Stephan / Löffler, Frank / Loewe, Axel / Renard, Bernhard Y. / Seemann, Gunnar / Struck, Alexander et al. (2021): “An environment for sustainable research software in Germany and beyond: current state, open challenges, and call for action [version 2; peer review: 2 approved].” *F1000Research* 9:295. https://doi.org/10.12688/f1000research.23224.2.

Barbot, Laure / Fischer, Frank / Moranville, Yoann / Pozdniakov, Ivan (2019). “Which DH Tools Are Actually Used in Research?” *weltliteratur.net. A Black Market for the Digital Humanities.* https://weltliteratur.net/dh-tools-used-in-research/ [last accessed: 17.04.2023].

Barbot, Laure / Fischer, Frank / Moranville, Yoann (2021). “ToolXtractor.” *GitHub.com.* https://github.com/lehkost/ToolXtractor [last accessed: 17.04.2023].

Chue Hong, Neil (ed.) (2019a): “Software Citation Checklist for Authors” (Version 0.9.0). *Zenodo.* http://doi.org/10.5281/zenodo.3479199.

Chue Hong, Neil (ed.) (2019b): "Software Citation Checklist for Developers" (Version 0.9.0). *Zenodo.* http://doi.org/10.5281/zenodo.3482769.

Druskat, Stephan (2021a): “Research software citation for researchers.” *Research Software Citation. Cite and Make Citable!* (Version 1.1). https://cite.research-software.org/researchers/ [last accessed: 04.11.2022].

Druskat, Stephan (2021b): “Research software citation for developers.” *Research Software Citation. Cite and Make Citable!* (Version 1.1). https://cite.research-software.org/developers/ [last accessed: 04.11.2022].

Henny-Krahmer, Ulrike / Jettka, Daniel (2021): “Softwarezitation in den Digital Humanities.” (Version 0.1). *Zenodo*. http://doi.org/10.5281/zenodo.5106.

Henny-Krahmer, Ulrike / Jettka, Daniel (2022): “Softwarezitation als Technik der Wissenschaftskultur. Vom Umgang mit Forschungssoftware in den Digital Humanities.” In: *DHd2022. Konferenzabstracts*. https://doi.org/10.5281/zenodo.6328046.

Hettrick, Simon / Antonioletti, Mario / Carr, Les / Chue Hong, Neil / Crouch, Stephen / De Roure, David / Emsley, Iain et al. (2014): “UK Research Software Survey 2014.” *Zenodo*. https://doi.org/10.5281/zenodo.14809.

Homburg, Timo / Klammt, Anne / Mara, Hubert / Schmid, Clemens / Schmidt, Sophie C. / Thiery, Florian / Trognitz, Martina (2020): “Diskussionsbeitrag - Handreichung zur Rezension von Forschungssoftware in den Altertumswissenschaften / Impulse - Recommendations for the review of archaeological research software.” *GitHub*. https://research-squirrel-engineers.github.io/Impuls_SoftwareRezensionen_DGUF/Draft.htm [last accessed: 04.11.2022].

Howison, James / Bullard, Julia (2016): “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 (JASIST)* 67 (9): 2137-2155. https://doi.org/10.1002/asi.23538.

Jackson, Mike (n. d.): “How to cite and describe software.” In: *Software and research: The Software Sustainability Institute’s Blog*. https://www.software.ac.uk/how-cite-and-describe-software [last accessed: 04.11.2021].

Jettka, Daniel / Henny-Krahmer, Ulrike (2022): “Leitfaden für die nachhaltige Entwicklung und Nutzung von Forschungssoftware.” *NFDI4Culture Handreichung* (Version 1.0.0). https://docs.nfdi4culture.de/ta3-sustainable-research-software [last accessed: 04.11.2022]. https://doi.org/10.5281/zenodo.7194401.

Jettka, Daniel / Henny-Krahmer, Ulrike / Ferger, Anne / Alvares Freire, Fernanda (2022): “ Software citation in the Digital Humanities.” (Version 0.1) GitHub.com. https://github.com/DH-RSE/software-citation [last accessed: 04.11.2022].

Krüger, Frank and David Schindler (2020): “A Literature Review on Methods for the Extraction of Usage Statements of Software and Data.” *Computing in Science & Engineering* 22 (1): 26-38. https://doi.org/10.1109/MCSE.2019.2943847.

Lamprecht, Anna-Lena / Garcia, Leyla / Kuzak, Mateusz / Martinez, Carlos / Arcila, Ricardo / Del Pico, Eva M. / Dominguez Del Angel, Victoria et al. (2020): “Towards FAIR principles for research software.” *Data Science* 3 (1): 37–59. https://doi.org/10.3233/DS-190026.

Smith, Arfon M. / Katz, Daniel S. / Niemeyer, Kyle E. / FORCE11 Software Citation Working Group (2016): “Software citation principles.” *PeerJ Computer Science* 2:e86. http://dx.doi.org/10.7717/peerj-cs.86.

Zarei, Alireza / Seung-Bin, Yim / Fischer, Frank /Ďurčo, Matej / Wieder, Phillipp (2022): “Measuring the Use of Tools and Software in the Digital Humanities: A Ma-chine-Learning Approach for Extracting Software Mentions from Scholarly Articles.” In: *DH 2022. Conference Abstracts* . https://dh2022.dhii.asia/abstracts/files/FISCHER_Frank_Measuring_the_Use_of_Tools_and_Software_in_the.html [last accessed: 04.11.2022].