Studying Processes in Software Citation Towards Improved Collaboration Among Scientists

Cai Fan Du
Johanna Cohoon
James Howison
cfdu@utexas.edu
jlcohoon@utexas.edu
jhowison@ischool.utexas.edu
The University of Texas at Austin
Austin, TX

Jason Priem
Heather Piwowar
jason@ourresearch.org
heather@ourresearch.org
Our Research Inc.
Vancouver, BC, Canada

ABSTRACT

We are seeking to understand the process of when and how the use of scientific software generates reputational rewards for scientific software contributors. At a more micro level, we are studying the practices surrounding software citation by researchers, scientific software producers, and publishers. Simultaneously, we are building a system called CiteAs, designed to improve the implementation and usefulness of software citation. Ultimately, we are interested in increasing collaboration in scientific software work. Our method is two-fold: we are conducting artifact-supported interviews and building a publicly available tool.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**;

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CSCW 2019 Workshop - Mapping the "How" of Collaborative Action, November 10, 2019, Austin TX, USA,

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

process, practice studies, empirical methods, scientific software, software citation

ACM Reference format:

Cai Fan Du, Johanna Cohoon, James Howison, Jason Priem, and Heather Piwowar. 2019. Studying Processes in Software Citation Towards Improved Collaboration Among Scientists. In *Proceedings of CSCW 2019 Workshop - Mapping the "How" of Collaborative Action,*
, November 10, 2019, Austin TX, USA, 7 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

WHY ARE THE PROCESSES IN SOFTWARE CITATION OF INTEREST

Software has been increasingly critical in scientific research. However, scientific software work has yet to fully find its place in research practice. Scientific work is dominated by a reputation economy based on the currency of scientific publications and citations. Scientific software work is often seen as "service work", staying largely invisible inside the system of citation and publication[2].

Earlier empirical work[2, 3] reveals that contributors to scientific software believed that due credit was not given and that visibility of their software contribution may not justify their time and effort spent on the building of software. Contributors held out hope that repurposing the system of citation and publication could improve the situation. By formally citing software in publications, researchers as software end-users would provide reputational rewards in a way that software producers could mobilize to demonstrate their impact[1].

In reality, the implementation of software citation is hard. Park and Wolfram [5] show that formal software citation practices are still relatively uncommon and that persistent identifiers like DOIs do not necessarily improve citation rates. Though scientific policy-makers, advocacy groups, and some publishers have been intervening to put software citation into action, challenges still exist. Such challenges include how to identify and cite software technically, and how to work with stakeholders to bring software citation into scholarly culture[4].

To implement software citation is to introduce behavioral change into the daily practice of researchers and other stakeholders in the scientific reward system. Therefore, we need to understand the existing practices surrounding software citation in order to identify entry points for effective intervention.

LOCALIZED PROCESSES THAT WE ARE TRYING TO RECONSTRUCT

We are seeking to understand the processes surrounding software citation at two different levels. At the macro level, we are trying to understand when and how the use of a piece of scientific software generates reputational reward for the contributors of the software. Each citation of the software used in research is an incremental reward that increases the visibility of software work. Such reward is expected to incentivize high-quality software work, especially through the collaboration among scientists. At a more micro level, we investigate the specific actions occurring at three key locations inside the system:

- Contributors making software available: how would they prefer to be cited or credited? How do they make this preference clear?
- Researchers authoring a paper: How do they come to the decision of what and how to cite? Can they find software authors' citation requests? Do they know how to format a useful software citation?
- The review and publication process of research papers: When and how are software citations reviewed? What citation guidelines are available?

HOW ARE WE STUDYING THESE PROCESSES?

Our approach is two-fold: First, we are conducting artifact-supported interviews for gaining an in-depth understanding of the localized processes; second, we have been developing a system called CiteAs to bridge the gaps in software citation [6]. In this way, we complement our research with a tool for intervention.

The artifact-supported interviews are designed for reconstructing our focal periods inside the system. The key is to reestablish critical moments from informant narratives, specifically discussing their past actions and underlying reasoning. We have been developing interview protocols tailored to each stakeholder role: end-user researchers, scientific software contributors, and academic publishers and journal reviewers. We seek to keep the interviews concrete by using artifact elicitation techniques. For end-user researchers, we identify their research publications and develop our conversation from these publications. For software contributors, we ask them to look for software citation requests online and tell us stories of attempts to ask for credit. At the moment, we are still constructing the interview protocol for academic publishers and journal reviewers; we intend to inquire about how they deal with citations in reviewing and publishing. Our interview questions also investigate the individual sociotechnical contexts, such as their understanding of how other researchers approach software citation, and their use of citation management software, etc.

Our system, CiteAs, has been used as an additional artifact for prompting further contextualized thinking from informants during interviews. We've asked informants to look for citation information via CiteAs during the interview while received their feedback on the system in development in the meantime. Primarily, CiteAs is a specialized search engine deliberately developed for seeking discoverable software citation requests online.

As a specialized search engine, CiteAs takes an input (e.g., name of a software package, link to a GitHub repository, DOI of a research product, etc.) and looks for requests for citation. The system follows a set of heuristics seeking the best guess of what the software authors would like to be cited: Package names are converted to related websites about the package by taking the first result from a search using the name string. Websites are spidered for identifying key phrases (e.g., "citation", "please cite", "citation.html", "citing.html") leading to a likely existing citation request. Links to GitHub repositories are followed and the repository files searched for specific citation request files (e.g., CITATION, CITATION.ccf, CODEMETA) or linked to language-specific citation files (e.g., R's DESCRIPTION file). DOIs can also be identified and the DOI API will be called to gather citation metadata. The system then returns its best guess of the citation desired by the software developers, formatted in a certain citation style. It also outputs the citation provenance information showing where that guess came from. Besides, we provide a way for users to report poor search results, and urge software contributors to improve the results by making their citation requests more clear and machine-readable.

THE CHALLENGES OF UNDERSTANDING SOFTWARE CITATION PROCESSES

The overall process of the reputational rewards as incentive feedback for scientific software development is vastly dispersed through time and geography. The systematic effect is substantial, but it arises from numerous individual actions across many locations at different times. Much of the relevant behaviors are both private and result in omissions from documents (not citing, or not requesting citation) and thus hard to be investigated. In our interviews, it is clear that many participants haven't actively thought about the topic, following familiar routines and unaware of the ultimate systematic effect. These issues are shared, we imagine, with many attempts to characterize system-level processes and phenomena in sociotechnical studies.

We understand there exists significant variation in the relevant practices of software citation across scientific fields, but as is often the case with research on sociotechnical systems we don't have a clear sampling frame. Currently, we are snowball sampling our participants and also try to diversify the roles, experience, and disciplinary areas of the recruited participants.

CURRENT STATUS OF THE PROJECT

Thus far we have conducted 5+ hour-long interviews with end-user researchers and 8+ hour-long interviews with scientific software contributors. Their professional roles, academic experience, and scientific fields vary greatly. We are still seeking more interview participants, especially publishers and journal editors to learn about their relevant practices.

Another ongoing endeavor of our project is that our team is manually annotating a dataset of software mentions in academic publications as the training set for a machine learning module of the system. It is our aim that CiteAs will be able to automatically identify informal software mentions in research papers and give citation suggestions. Such efforts, combined with the search capability of

CiteAs, are expected to improve the implementation and usefulness of software citation in reaction to the existing challenges. Our full conception of CiteAs is illustrated in Figure 1.

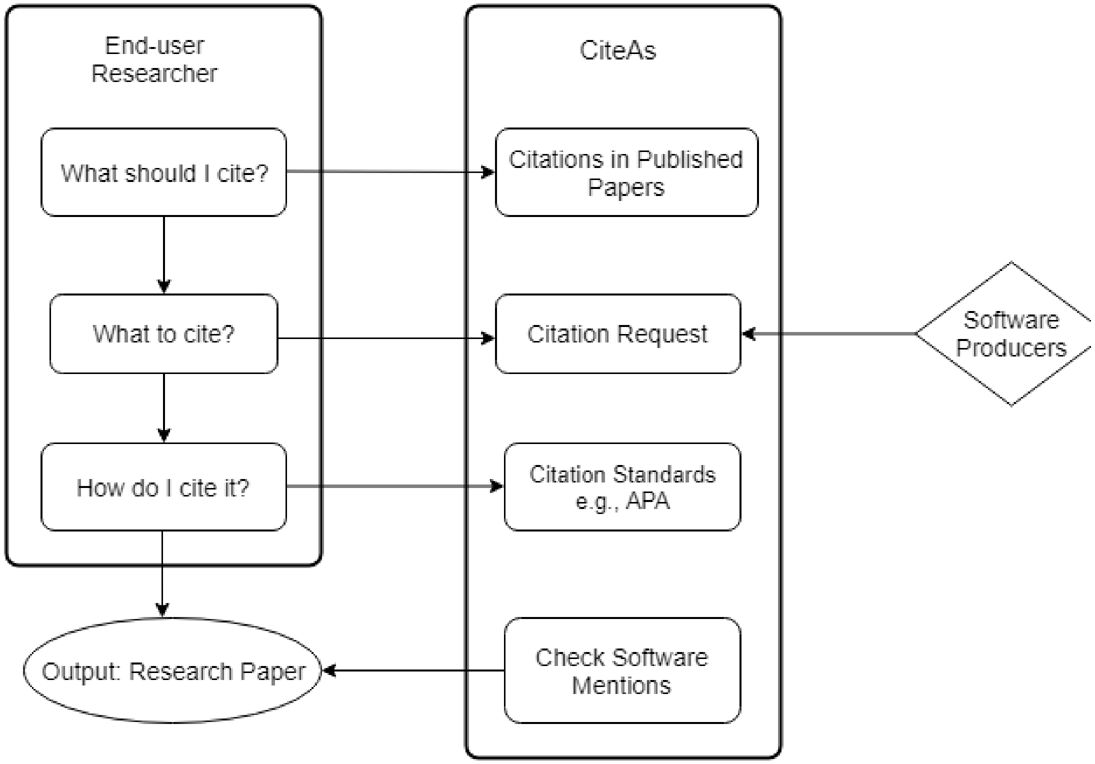

Figure 1: CiteAs is designed for bridging the gaps in software citation

The full development of CiteAs will realize 1) automatic detection of informal software mentions in published papers, informing end-users what they should cite in their work; 2) identify citation requests made by software producers and pass on this information to end-user researchers; 3) recommend software citation in a chosen citation format to standardize end-users' citation practice.

ACKNOWLEDGMENTS

This material is based upon work supported by the Sloan Foundation's Digital Science program.

REFERENCES

[1] James Howison, Ewa Deelman, Michael J McLennan, Rafael Ferreira da Silva, and James D Herbsleb. 2015. Understanding the scientific software ecosystem and its impact: Current and future measures. *Research Evaluation* 24, 4 (2015), 454–470.

[2] James Howison and James D Herbsleb. 2011. Scientific software production: incentives and collaboration. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 513–522.

[3] James Howison and James D Herbsleb. 2013. Incentives and integration in scientific software production. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 459–470.

[4] Daniel S Katz, Daina Bouquin, Neil P Chue Hong, Jessica Hausman, Catherine Jones, Daniel Chivvis, Tim Clark, Mercè Crosas, Stephan Druskat, Martin Fenner, et al. 2019. Software Citation Implementation Challenges. *arXiv preprint arXiv:1905.08674* (2019).

[5] Hyoungjoo Park and Dietmar Wolfram. 2019. Research software citation in the Data Citation Index: Current practices and implications for research software sharing and reuse. *Journal of Informetrics* 13, 2 (2019), 574–582.

[6] Jason Priem, Heather Piwowar, and James Howison. 2019. CiteAs. (2019). http://citeas.org

BIBLIOGRAPHY SKETCHES

Cai Fan Du

Fan is a doctoral student at the Information School of the University of Texas at Austin. She studies the production of digital artifacts, especially in the context of scientific research. Her vision is sociotechnical, with a focus on the organizing of distributed work and the institutional change it might entail. Before starting her doctoral program, she did research on digital technology adoption and use to inform local government's policy-making and worked in system engineering project in industry. she holds a B. in Economics and a B.A. in french.

Johanna Cohoon

Johanna (Hannah) Cohoon is a PhD Candidate at the Information School of the University of Texas at Austin where she studies scientific research practices and cyberinfrastructure. Previously, she worked at the Center for Open Science where she researched reproducibility in Psychology. She is interested in how scientific norms and practices change. Johanna received her bachelors degree in Cognitive Science from the University of Virginia.

James Howison

James Howison is an Associate Professor at the Information School of the University of Texas at Austin, where he has been since August 2011, following a post-doc at CMU and his 2009 PhD from Syracuse Information School. James has studied open source software development and the development of software in science because both are interesting examples of collaboration, and is particularly interested in understanding how different incentives, such as working for fun or for academic reputation, lead to different structures of collaboration. James is a 2019 PECASE award winner, based on a NSF 2014 CAREER award. Publications include MISQ, Information and Organization, and CSCW. http://james.howison.name