A Retrospective Study of one Decade of Artifact Evaluations

Stefan Winter [1], Christopher Timperley [2], Ben Hermann [3], Jürgen Cito [4], Jonathan Bell [5], Michael Hilton [2], and Dirk Beyer [1]

Abstract: Reproducibility is a vital property of experimental and empirical research, without which it is difficult to establish trust in derived conclusions. If results cannot be independently confirmed, they may be affected by observer bias or other confounding factors. As the full-scale reproduction of scientific results from a study takes significant time, which does not match well with the conference-focused publication in computer science, a lighter quality assurance mechanism for scientific work has been established. Starting with ESEC/FSE in 2011, the software engineering and programming languages communities have increasingly adopted *artifact evaluations* to assess research artifacts, with which the submitted results have been obtained, as a light-weight mechanism to increase trust in the results.

After a decade of artifact evaluations, we analyze the impact they have had on published articles and artifacts. Our findings show that the artifact publication standards, which have improved with artifact evaluations, have had a positive effect on the availability of published artifacts. At the same time, they point out avenues for further improving artifact availability and documentation as well as the visibility of the associated publications. Our original publication of the study has been presented at FSE 2022 [Wi22].

Keywords: Metascience, Empirical software engineering, Mining software repositories

Study Overview and Results Summary

In the software engineering and programming languages communities, artifact evaluation (AE) has been introduced at ESEC/FSE 2011 with OOPSLA and ECOOP following in 2013. Since then, it has become a commonplace activity at top ranked conferences with numerous submissions reviewed by large artifact evaluation committees. With more than a decade of AE data and insights, we have conducted a study to assess if AE is effective and to identify improvement opportunities for the coming years [Wi22]. Contrary to prior research in the field, our study solely focuses on properties of the published articles and artifacts (as opposed to survey-based [HWS20] or survey-supported [Ti21] work) and on the effect of artifact evaluations (as opposed to how unevaluated artifacts are shared [CP16, He20]).

1 LMU Munich, Munich, Germany, sw@stefan-winter.net, https://orcid.org/0000-0001-8244-995X;
dirk.beyer@sosy-lab.org, https://orcid.org/0000-0003-4832-7662
2 Carnegie Mellon University, Pittsburgh, PA, USA,
ctimperley@cmu.edu, https://orcid.org/0000-0002-9785-324X;
mhilton@cmu.edu, https://orcid.org/0000-0001-9195-6902
3 TU Dortmund, Dortmund, Germany,
ben.hermann@cs.tu-dortmund.de, https://orcid.org/0000-0001-9848-2017
4 TU Wien, Wien, Austria, juergen.cito@tuwien.ac.at, https://orcid.org/0000-0001-8619-1271
5 North Eastern University, Boston, MA, USA,
j.bell@northeastern.edu, https://orcid.org/0000-0002-1187-9298

doi:10.18420/sw2024_28

To this end, we have analyzed 3 650 research articles from 64 proceedings of 12 software engineering and programming languages conferences. We checked if they successfully passed AE and answered the following research questions on that basis.

RQ1 Are articles with artifacts that have passed AE more visible?
We have compared citations across the AE and non-AE article groups and find no statistically significant differences after controlling for confounding factors (paper length, open/closed access). This suggests that alternative reward mechanisms are need for high-quality artifacts.

RQ2 Are successfully evaluated artifacts more available?
We find that successful retrieval of artifacts is strongly linked with "Available" badges on articles. As the criteria for this badge are linked to artifact *publication aspects*, rather than if/how they have been evaluated in AE, we suggest to advertise the badge in calls for papers, rather than calls for artifacts, and check them on the publisher side.

RQ3 Is artifact development/maintenance continued more often for successfully evaluated artifacts?
We find stronger indications of visibility, interest, and continued artifact development for evaluated artifacts on GitHub. Not evaluated artifacts seem to use GitHub as a free storage solution (for which better alternatives exist).

RQ4 Are successfully evaluated artifacts more often reused?
There is no clear indication of an advantage for AE, similar to the findings for RQ1: Within our dataset for the study, we find that more AE articles are cited than non-AE papers. However, the number of references to the non-AE articles are higher.

RQ5 Are successfully evaluated artifacts more thoroughly documented?
In a sample of 100 AE and non-AE artifacts each, we find the majority to be documented. In both cases, more than 10 % had no documentation and about 50 % had no licenses. We hope this to improve with the documentation standards introduced at ESEC/FSE 2018, which are increasingly adopted by other conferences.

Bibliography

[CP16] Collberg, Christian; Proebsting, Todd A.: Repeatability in Computer Systems Research. Commun. ACM, 59(3):62–69, feb 2016.

[He20] Heumüller, Robert; Nielebock, Sebastian; Krüger, Jacob; Ortmeier, Frank: Publish or perish, but do not forget your software artifacts. Empir. Softw. Eng., 25(6):4585–4616, 2020.

[HWS20] Hermann, Ben; Winter, Stefan; Siegmund, Janet: Community Expectations for Research Artifacts and Evaluation Processes. In: Proc. ESEC/FSE. ACM, pp. 469–480, 2020.

[Ti21] Timperley, Christopher Steven; Herckis, Lauren; Goues, Claire Le; Hilton, Michael: Understanding and Improving Artifact Sharing in Software Engineering Research. Empir. Softw. Eng., 26(4):67, 2021.

[Wi22] Winter, Stefan; Timperley, Christopher S.; Hermann, Ben; Cito, Jürgen; Bell, Jonathan; Hilton, Michael; Beyer, Dirk: A Retrospective Study of One Decade of Artifact Evaluations. In: Proc. ESEC/FSE. ACM, p. 145–156, 2022.