Artifact Evaluations for Stronger Research Results

Dirk Beyer
dirk.beyer@sosy.ifi.lmu.de
LMU Munich
Germany

Stefan Winter
sw@stefan-winter.net
LMU Munich
Germany

Abstract

Over the past decades, reports of *reproducibility crises* have surfaced in various scientific communities. Independent confirmations of published research results failed, casting doubt on the validity of these results. Even before the magnitude of the problem has become apparent in many domains, the software-engineering community introduced artifact evaluations, for the first time at ESEC/FSE 2011, in which research artifacts that support published results were voluntarily submitted for peer review. Since then, artifact evaluations have become immensely popular and are today being offered to authors at most software-engineering venues, where large artifact-evaluation committees handle large numbers of artifact submissions. At some venues, papers are accepted for publication only if their artifacts pass the artifact evaluation.

To make sure that this enormous and important effort from our community to (a) create and (b) assess research artifacts is well-spent, knowledge and insights from successful and unsuccessful artifact-evaluation practices as well as publishing implications need to be conserved and shared with prospective participants, i.e., authors, reviewers, and organizers. Based on insights from empirical studies about artifact evaluations in the software-engineering community, from running artifact evaluations at different conferences, and from managing publication processes after artifact acceptance, this tutorial presents an overview what artifact evaluations are and how they are conducted, along with known pitfalls and established best practices to overcome them. The presented insights will be accompanied by a hands-on training session on artifact evaluation using published research artifacts. The tutorial targets prospective artifact-evaluation organizers and reviewers as well as researchers wishing to strengthen their research results through the research artifacts they create.

Keywords

Research Artifacts, Artifact Evaluations, Reproducibility

ACM Reference Format:
Dirk Beyer and Stefan Winter. 2025. Artifact Evaluations for Stronger Research Results. In *33rd ACM International Conference on the Foundations of Software Engineering (FSE Companion '25), June 23–28, 2025, Trondheim, Norway*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3696630.3728623

FSE Companion '25, Trondheim, Norway

ACM ISBN 979-8-4007-1276-0/2025/06
https://doi.org/10.1145/3696630.3728623

1 Tutorial Aims and Objectives

The tutorial aims to

- improve artifact-evaluation practices and
- support researchers who are creating research artifacts

by sharing knowledge about pitfalls and best practices, and by providing practical examples and hands-on experience with artifact evaluations.

The provided insights mainly result from experiences with artifacts in software-engineering and programming-languages research. Researchers from these domains are, thus, expected to benefit most from the tutorial. At the same time, many of the insights are pertaining to the broader ACM artifact badge definitions and to artifact packaging and publication, which are not limited to any specific domain. These broader insights apply to any type of research that can be supported by any type of artifact; code, data records, mechanized proofs, etc. The tutorial will also discuss examples for which artifacts cannot be published or for which they are of lesser utility, and how such cases can nonetheless benefit from artifact evaluations, albeit to a lesser degree.

As the tutorial aims to contribute to the community knowledge of better and worse practices, we will discuss decisions that artifact authors, reviewers, or artifact-evaluation organizers have made and that turned out to have undesirable consequences in hindsight. To learn from these problems, it is unnecessary to link them with individual researchers. We will, hence, not make any such link or support tutorial participants in making such links.

2 Intended Audience and Required Background

The tutorial targets researchers at all career levels as authors, reviewers, and conference organizers. It does not require prior knowledge or experience with artifact evaluations. As most published research artifacts are created for GNU/Linux-based systems, basic knowledge how to operate GNU/Linux systems is beneficial for the hands-on part of the tutorial.

3 Relevance

Artifact evaluations have become a wide-spread activity at software-engineering venues with increasingly large committees. As these committees are often assembled from junior researchers, the churn is usually higher than for program committees. Moreover, a study presented at ESEC/FSE 2020 has found that a large fraction of artifact-evaluation committee (AEC) members has served on the committee with no prior experience creating artifacts [2]. We also experience regularly that reviewers and organizers of artifact evaluations are often missing information about the implications of the badge definitions, which leads to (unintended) inconsistencies between the badge policies defined by the publishers and how they are implemented at conferences. Finally, as presented in detail in a study

at ESEC/FSE 2022, we often see artifact publications that are quickly rendered unusable [4], even within a short time after publication.

Considering the effort that usually goes into artifact creation and evaluation, we consider this undesirable. We hope that dedicated training sessions for artifact creation and evaluations will help to improve the situation.

4 Format

The tutorial will comprise two parts, a lecture to provide knowledge and facts, and a hands-on session to provide participants with the opportunity to experience the practical implications.

5 Intended Duration

Each of the two parts is planned for 90 minutes, i.e., 3 hours in total.

6 Outline of Topics

6.1 Overview

The covered topics include

- an introduction and historical perspective on reproducibility and artifact evaluations,
- an overview of ACM's policy on artifacts and reproducibility,
- a discussion of common problems encountered when (re-)using published artifacts and solutions for those problems covering actionable advise for authors, reviewers, and conference organizers,
- a time line and typically provided documents as templates for artifact-evaluation chairs,
- recommendations for conducting artifact reviews,
- a guideline for preparing a reproducibility artifact, and
- a hands-on session in which a few example artifacts are evaluated by the audience.

As a reference for tutorial participants, we provide a brief description of the relevant concepts and terminology as well as a proposed artifact-evaluation guideline in Sections 6.2 and 6.3.

6.2 Artifact Badging and ACM Policy

In the software-engineering community, artifact evaluations have started as a grassroots effort at ESEC/FSE 2011[1] and were soon adopted by other conferences. At OOPSLA 2013, Matthias Hauswirth and Steve Blackburn introduced a badge[2] to indicate successful artifact evaluations on papers and thereby give successfully evaluated artifacts more visibility (see Figure 1). In 2017, ACM presented a set of artifact evaluation badges along with definitions for a fundamental terminology for artifact evaluations and their intended contribution[3]. According to this terminology, an artifact is defined as "a digital object that was either created by the authors to be used as part of the study or generated by the experiment itself". ACM also organized a community task-force workshop on defining important principles [1]. In 2020, ACM's set of badges was slightly revised[4]

Figure 1: Badge for OOPSLA 2013

Figure 2: ACM badges v1.1 (since 2020)

to better align their underlying terminology with the NISO recommended practice for reproducibility badging and definitions [3]. This set of badges is currently used by a large number of conferences and has even been adopted by non-ACM organizations, such as EAPLS[5] and ETAPS[6], with ACM's permission. The current badges are shown in Figure 2 and fall into three categories indicated by their different primary colors in Figure 2. The following descriptions summarize, simplify, and partially interpret the definitions.

- **Artifacts Available (green):** The artifacts have been published on a publicly accessible archival repository with a declared plan to enable permanent accessibility (usually, a DOI is required).
- **Artifacts Evaluated (red):** The artifacts have successfully undergone an independent audit (the artifact evaluation). Two levels are distinguished:
  - **Functional**: The artifact can be used to support claims in the paper.
  - **Reusable:** The artifact is well structured and documented, so that it is expected to support claims from the paper for an extended period of time and may be repurposed beyond the paper's scope.

  Reusable implies *Functional* and only one of the two should be awarded.
- **Results Validated (blue):** Badges in this category focus on the validation of research results.
  - **Results Reproduced**: The results from the paper are confirmed in an independent study using artifacts from the original paper.
  - **Results Replicated**: The results from the paper are confirmed in an independent study *without* using artifacts from the original paper.

  As the blue "Results Validated" badges are pertaining to results presented in the paper, rather than properties of a related artifact, they are usually not relevant for artifact evaluations.

6.3 Review Guidelines

Over the years, ACM's definitions outlined in Section 6.2 have been interpreted differently by artifact-evaluation chairs [2]. Based on

[1] http://2011.esec-fse.org/cfp-artifact-evaluation, accessed 2025-04-13
[2] http://evaluate.inf.usi.ch/artifacts/aea/badge.html, accessed 2025-04-13
[3] https://www.acm.org/publications/policies/artifact-review-badging, accessed 2025-04-13
[4] https://www.acm.org/publications/policies/artifact-review-and-badging-current, accessed 2025-04-13
[5] https://eapls.org/pages/artifact_badges/, accessed 2025-04-13
[6] https://etaps.org/about/artifact-badges/, accessed 2025-04-13

community expectations expressed in a survey among artifact-evaluation committee members, on discussions with artifact-evaluation chairs, and on our experience with running artifact evaluations at different conferences, we have come up with review check-lists to help artifact reviewers with their badging decisions. In accordance with existing practices, we recommend to split the artifact review in two phases, a pre-assessment ("kick-the-tires") review to check that the artifact is operable and a full review to make a badging decision for operable artifacts. For the two phases, we offer the following questions as guidance to artifact reviewers and participants in the tutorial.

Pre-assessment ("Kick the tires"):

- ☐ Can the digital object be retrieved without revealing your identity?
- ☐ Is it packaged in an open format that you can work with?
- ☐ If the provided digital object is a compressed archive, can it be decompressed without errors?
- ☐ Is the artifact operable on your computing architecture?
- ☐ Are all documents required in the call for artifacts included in the submission?
- ☐ Do the documents contain the information asked for in the call for artifacts?

Full Review (Badging Decisions):

- **Available:**
  - ☐ Is the artifact published on a long-term archival platform with declared retention policy (≥ 10 years)?
  - ☐ Is a DOI provided? (A DOI implies long-term availability.)
  - ☐ If the artifact is on Zenodo: Is it linked to via a version-specific DOI (as opposed to a concept DOI, which is always redirected to the latest version and should be avoided to support reproducibility)?
  - ☐ Is the artifact linked to from the paper using the DOI link (or other long-term archive link)?
  - ☐ Is a license specified for the artifact?
- **Functional:**
  - ☐ Is the artifact exercisable?
  - ☐ Have the documentation guidelines in the call for artifacts been followed?
  - ☐ Is the artifact sufficiently documented to be used for reproducing results from the paper?
  - ☐ Does the artifact contain all relevant parts for reproducing results from the paper (are input data, plotting scripts, etc. included; are external dependencies expected to be permanently available)?
  - ☐ Are results you have obtained using the artifact consistent with what is written in the paper?
- **Reusable:**
  - ☐ Is the artifact documentation sufficiently comprehensive and well structured, such that the artifact can be used in other settings than reproducing the exact study presented in the paper?
  - ☐ Are common standards for code, mechanized proofs, and data formats followed, so that it can be adapted with reasonable effort?

7 Key Learning Objectives

After attending the tutorial, participants are expected to have a thorough (conceptual and practical) understanding

- of the goals of artifact evaluations and how they are conducted,
- of the existing artifact-badging systems, the requirements they entail, and related terminology,
- of publishing aspects, such as licensing and existing options for long-term archiving,
- of the common threats to reproducibility and reuse in published artifacts,
- how submission requirements and review guidelines can support better artifact publications,
- how interactions across the PC and AEC chairs can narrow the gap between reviewed and published artifacts, and
- what authors can do to facilitate artifact reviews and ensure long-time utility of their artifacts.

8 Presenter's Bio

Dirk Beyer is a full professor for computer science at LMU Munich, Germany. Dirk is the co-author of more than 100 artifacts, has served on the ACM Task Force on Data, Software, and Reproducibility in Publication[7] [1], and has been actively shaping artifact publication practices through his community service for the ETAPS conferences[8], as well as through the publication services he has been offering to various software-engineering and programming-languages conferences. He is chairing the award committee for the Rance Cleaveland Test-of-Time Tool Award[9] and has co-authored an empirical study on artifact-evaluation practices presented at ESEC/FSE 2022 [4].

Stefan Winter is a postdoctoral researcher at LMU Munich, Germany, where he regularly offers seminars on reproducibility in software-engineering research. He has previously been teaching a course on reproducibility of software-based measurements at Ulm University. Stefan has been chairing the artifact-evaluation committees for ECOOP (2022, 2023) and FASE (2024, 2025) and has been providing advice on artifact-evaluation processes to conference organizers. He is a member of the Rance Cleaveland Test-of-Time Tool Award[9] committee and has co-authored two papers on artifact evaluations presented at ESEC/FSE in 2020 [2] and 2022 [4].

9 Tutorial History

None.

10 Audio-Visual and Technical Requirements

For the presentation in the first part of the tutorial, a projector, speakers, and microphones for presenter and audience are required. A fast and stable Internet connection is required for the hands-on part. If the latter cannot be provided reliably, a WiFi router and storage system with around 500 GB capacity can serve as a replacement.

[7] https://www.acm.org/publications/task-force-on-data-software-and-reproducibility, accessed 2025-04-13
[8] https://tacas.info/artifacts-best-practices.php, accessed 2025-04-13
[9] https://etaps.org/awards/test-of-time-tool/, accessed 2025-04-13

References

[1] Simon Adar, Dirk Beyer, Patricia Cruse, Gustavo Durand, Wayne Graves, Christopher Heid, Lundon Holmes, Chuck Koscher, Meredith Morovatis, Joshua Pyle, Bernard Rous, Wes Royer, and Dan Valen. 2017. Best Practices on Artifact Integration. Zenodo. doi:10.5281/zenodo.7296608

[2] Ben Hermann, Stefan Winter, and Janet Siegmund. 2020. Community expectations for research artifacts and evaluation processes. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Virtual Event, USA) *(ESEC/FSE 2020)*. Association for Computing Machinery, New York, NY, USA, 469–480. ISBN: 9781450370431 doi:10.1145/3368089.3409767

[3] NISO. 2021. *Reproducibility Badging and Definitions: A Recommended Practice of the National Information Standards Organization.* Technical Report NISO RP-31-2021. doi:10.3789/niso-rp-31-2021

[4] Stefan Winter, Christopher S. Timperley, Ben Hermann, Jürgen Cito, Jonathan Bell, Michael Hilton, and Dirk Beyer. 2022. A retrospective study of one decade of artifact evaluations. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Singapore, Singapore) *(ESEC/FSE 2022)*. Association for Computing Machinery, New York, NY, USA, 145–156. ISBN: 9781450394130 doi:10.1145/3540250.3549172