



# Community Expectations for Research Artifacts and Evaluation Processes

Ben Hermann, Stefan Winter, Janet Siegmund

technische universität dortmund

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

TECHNISCHE UNIVERSITÄT CHEMNITZ

Paper at ESEC/FSE 2020, Presentation at GI SE’23

ACM SIGSOFT Distinguished Paper Award

# Research Artifact

*… a self-contained work result with a context-specific purpose.*

Daniel Méndez Fernández, Wolfgang Böhm, Andreas Vogelsang, Jakob Mund, Manfred Broy, Marco Kuhrmann, and Thorsten Weyer. 2019. Artefacts in software engineering: a fundamental positioning. Software & Systems Modeling 18, 5 (2019), 2777–2786

# Reproducibility of Research is Complicated

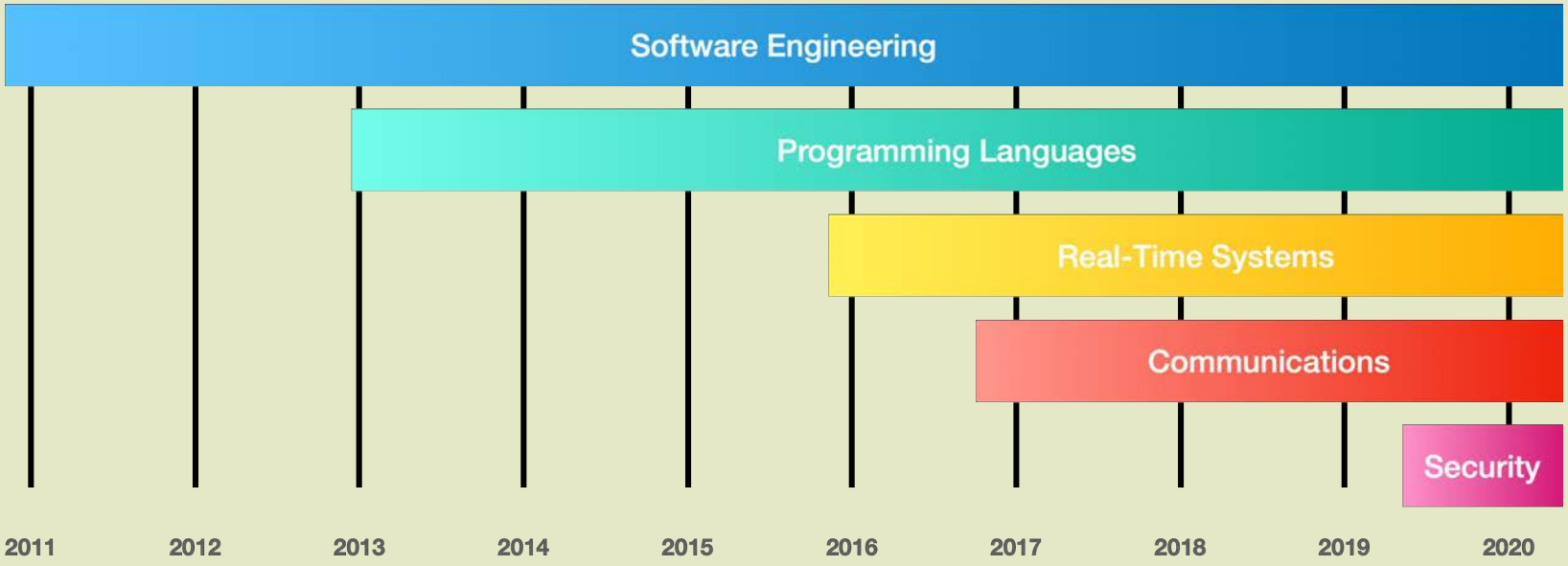
Numerous studies in multiple fields (including CS) report obstacles or failure of reproducing other researchers results

Oftentimes prototypes or data are not available

Prototypes oftentimes do not run (anymore) or have incomplete documentation

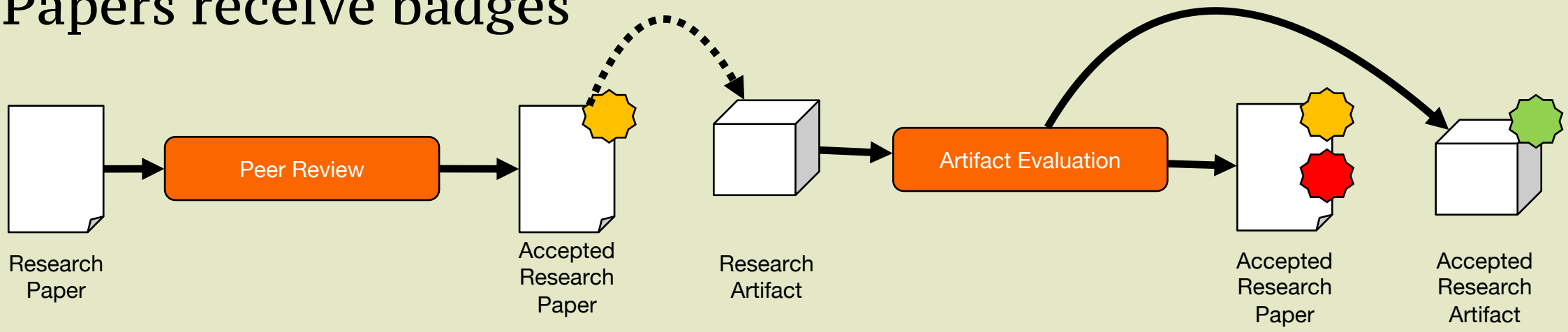
**Reproduction as a means of knowledge building and validation is impossible to achieve this way**

# Artifact Evaluation to the Rescue

# Artifact Evaluation: Quick Overview

- (At most venues) artifact evaluation is a voluntary process
- Meaning it does not influence paper acceptance (yet)
- Artifacts are peer reviewed by a different committee than the paper
- Papers receive badges



# Does Artifact Evalution Foster Better Artifacts?

But what is a “better” artifact?

# Our Study

Goals for our qualitative study

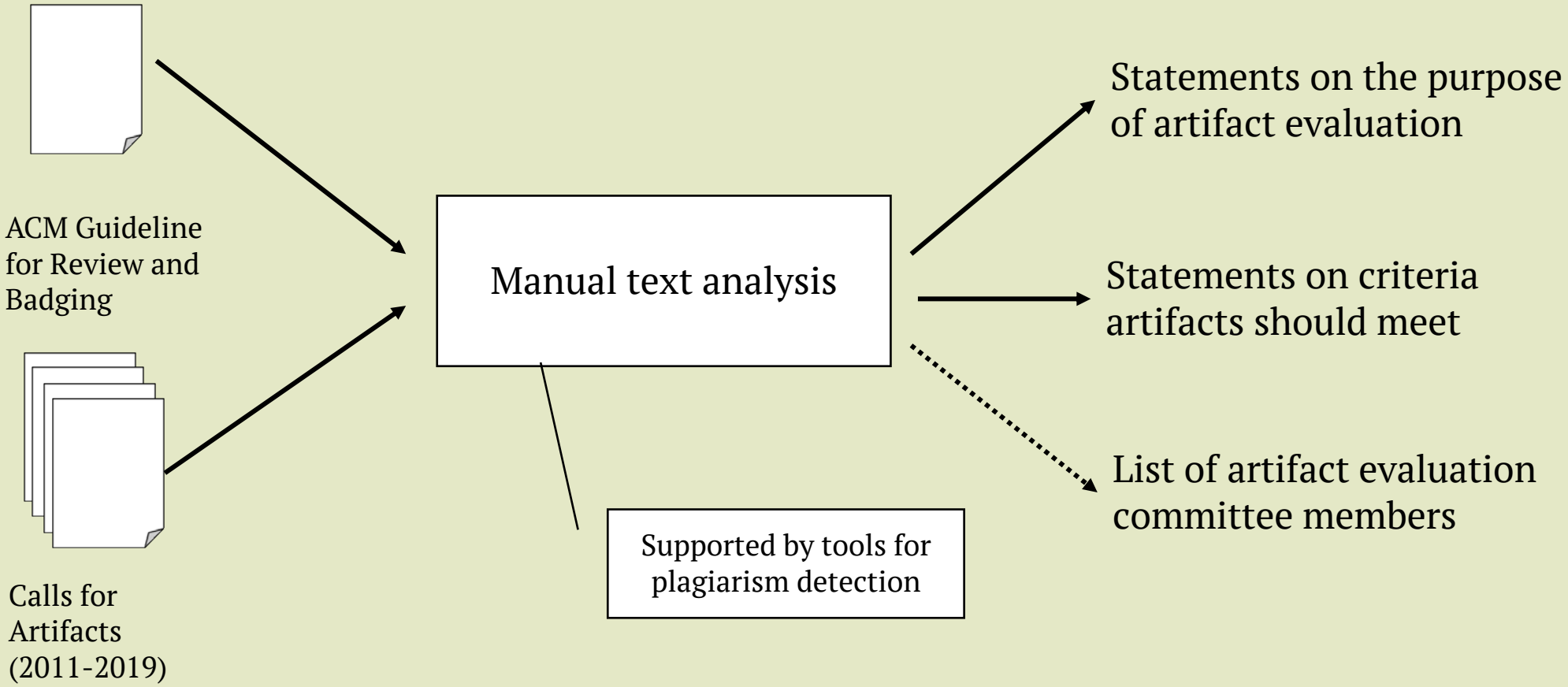
Find the perceived **purpose of artifact evaluation**

**Quality criteria** and **expectations** for research artifacts

**Difference** between the Software Engineering and Programming Language Community

# Methodology

## Pre-Study

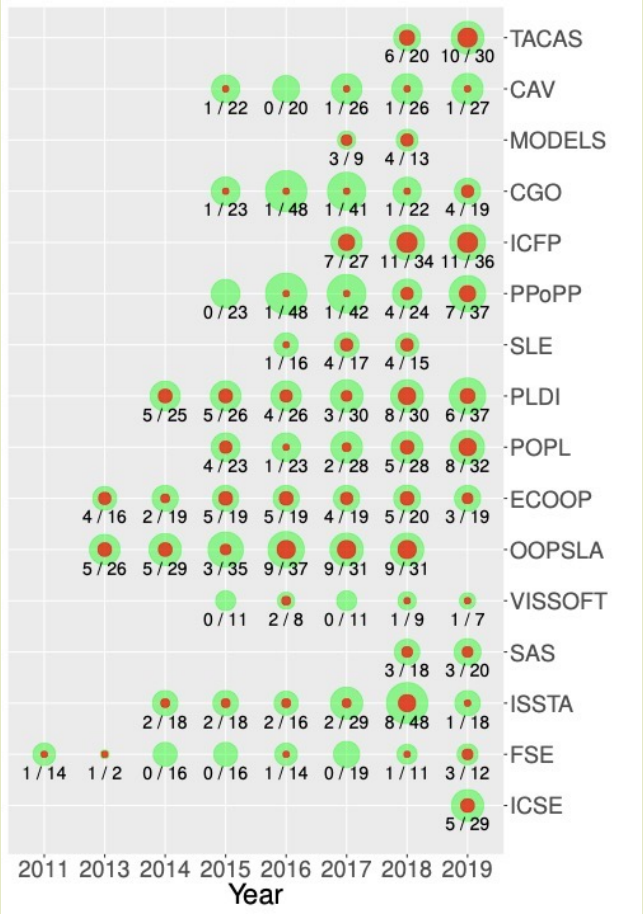


# Methodology

Online Survey

- All members of AECs (1,034 individuals) in SE and PL
- Goal: Discover the importance of review criteria and the purpose of AE
- Received 257 responses (≈ 25% response rate)
  - 124 complete, 133 incomplete
  - both included in the study
- Open answers
- Answers were coded using Hudson's open card sorting method: one researcher labelled answers, second researcher checked labels, differences were discussed until consensus was reached

# Inspected Conferences and Years




Invited (green), Responses Received (red)

We selected venues from the Software Engineering and Programming Language Field that perform Artifact Evaluation

We received answers for almost all conferences and years

# RQ1: Purpose of Artifact Evaluation

- Reuse of Artifacts
  - More mentioned in CfAs in SE than in PL
- Reproducibility of Results
  - More mentioned in CfAs in PL than in SE
  - AE should validate claims
  - AE should validate results
- Replicability and Reproducibility often confused
  - Inconsistent use in the community

# Repeatability, Reproducibility, Replicability

| Repeatability | Reproducibility | Replicability |
| --- | --- | --- |
| Original Team | Different Team | Different Team |
| Original Setup | Original Setup | Different Setup |

https://www.acm.org/publications/policies/artifact-review-and-badging-current

# Dual Purpose

Reuse ⟷ Replicability

These goals have **different implications** for artifact creation and review

In our survey, reproducibility was **not viewed as a beneficial factor** for reusability

**Less than half** of the respondents report experience with artifact reuse

# RQ2: Quality Expectations

**No consensus** of a quality threshold

ACM Guideline gives rough dimensions, but leaves definition of the declared dimensions open

Artifact types (code, data, mechanized proofs) differ in expectations

- Code and data artifacts are expected to have proper documentation
- For mechanized proofs completeness and understandability dominated the answers

Standardizations appears to lower the burden of portability

# Further Insights

Mixed satisfaction with the process

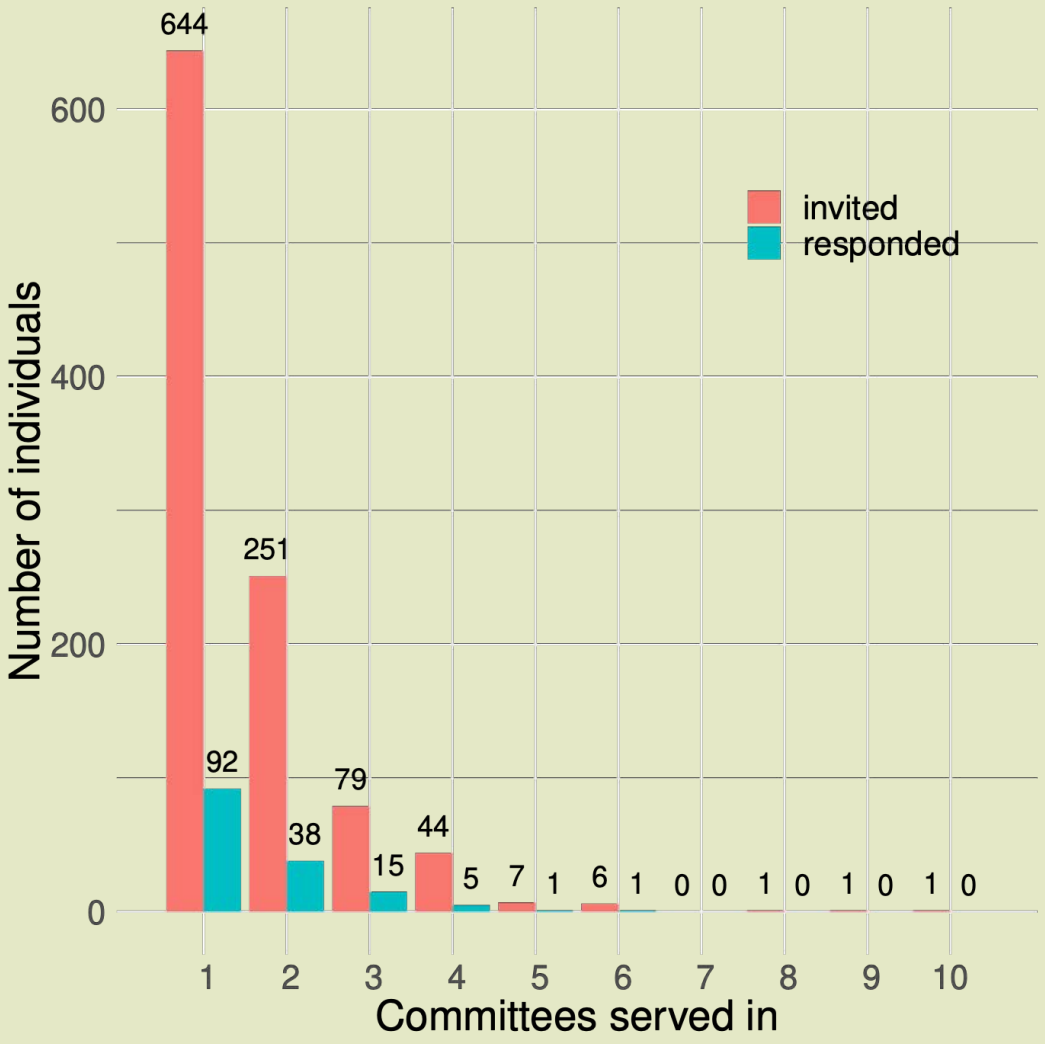
More criteria and guidelines needed

Only 66% of respondents had experience creating an artifact

**62% of artifact reviewers only serve once (cf. diagram)**

Review as an interactive process

Tighter coupling to paper acceptance



# Suggestions

Reproducibility != Reusability… What to evaluate for?

- ➔(Chairs) Define a primary goal and communicate explicitly

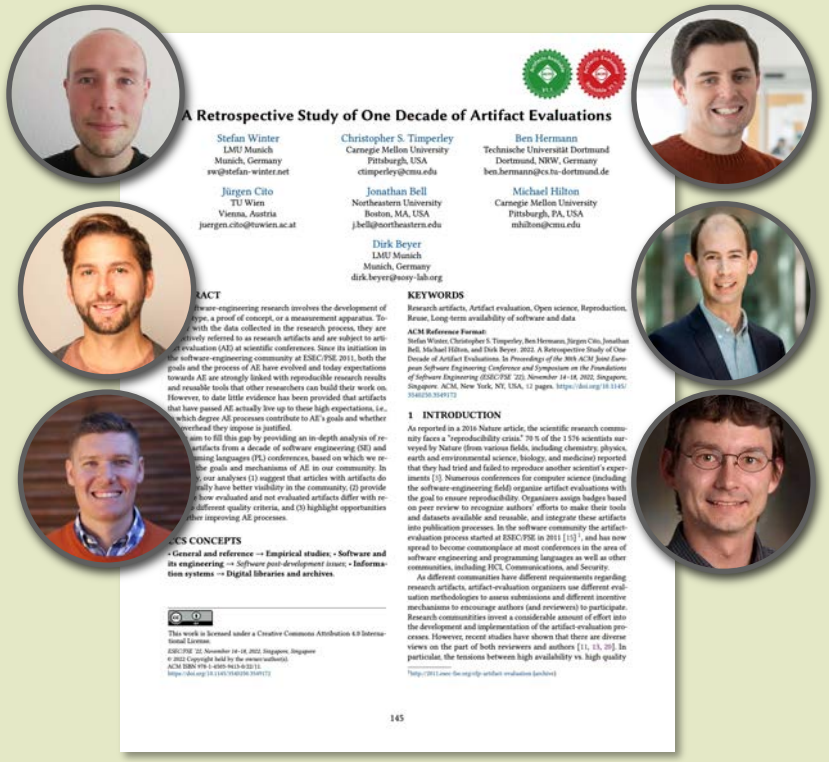
Ambiguous quality criteria for artifacts & artifact types

- ➔(Chairs and Steering Committees) Badges should have the same meaning between conferences and years
- ➔(Chairs and Steering Committees) Establish unified quality standards

Reviewers may be unexperienced

- ➔(Chairs) Recruiting criteria should take this into account
- ➔(Chairs) Provide detailed review guidelines

# Follow-up Work



A Retrospective Study of One Decade of Artifact Evaluations

Stefan Winter, LMU Munich, Munich, Germany, sw@stefan-winter.net
Christopher S. Timperley, Carnegie Mellon University, Pittsburgh, USA, ctimperley@cmu.edu
Ben Hermann, Technische Universität Dortmund, Dortmund, NRW, Germany, ben.hermann@cs.tu-dortmund.de
Jürgen Cito, TU Wien, Vienna, Austria, juergen.cito@tuwien.ac.at
Jonathan Bell, Northeastern University, Boston, MA, USA, j.bell@northeastern.edu
Michael Hilton, Carnegie Mellon University, Pittsburgh, PA, USA, mhilton@cmu.edu
Dirk Beyer, LMU Munich, Munich, Germany, dirk.beyer@sosy-lab.org

KEYWORDS
Research artifacts, Artifact evaluation, Open science, Reproduction, Reuse, Long-term availability of software and data

ACM Reference Format:
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 (ESEC/FSE '22), November 14–18, 2022, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3540250.3549172

1 INTRODUCTION

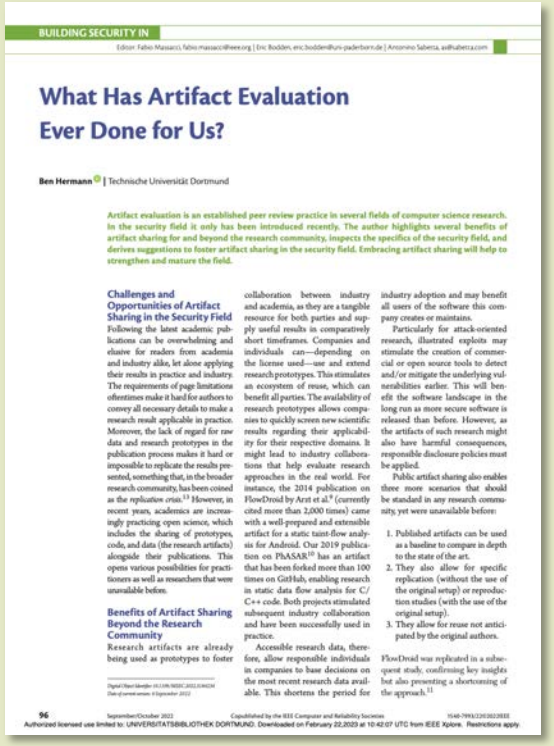
As reported in a 2016 Nature article, the scientific research community faces a "reproducibility crisis." 70 % of the 1 576 scientists surveyed by Nature (from various fields, including chemistry, physics, earth and environmental science, biology, and medicine) reported that they had tried and failed to reproduce another scientist's experiments [1]. Numerous conferences for computer science (including the software-engineering field) organize artifact evaluations with the goal to ensure reproducibility. Organizers assign badges based on peer review to recognize authors' efforts to make their tools and datasets available and reusable, and integrate these artifacts into publication processes. In the software community the artifact-evaluation process started at ESEC/FSE in 2011 [15][1], and has now spread to become commonplace at most conferences in the area of software engineering and programming languages as well as other communities, including HCI, Communications, and Security.

As different communities have different requirements regarding research artifacts, artifact-evaluation organizers use different evaluation methodologies to assess submissions and different incentive mechanisms to encourage authors (and reviewers) to participate. Research communities invest a considerable amount of effort into the development and implementation of the artifact-evaluation processes. However, recent studies have shown that there are diverse views on the part of both reviewers and authors [11, 13, 20]. In particular, the tensions between high availability vs. high quality





S. Winter, C. S. Timperley, B. Hermann, J. Cito, J. Bell, M. Hilton, and D. 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 (ESEC/FSE 2022). https://doi.org/10.1145/3540250.3549172



BUILDING SECURITY IN

What Has Artifact Evaluation Ever Done for Us?

Ben Hermann | Technische Universität Dortmund

Artifact evaluation is an established peer review practice in several fields of computer science research. In the security field it only has been introduced recently. The author highlights several benefits of artifact sharing for and beyond the research community, inspects the specifics of the security field, and derives suggestions to foster artifact sharing in the security field. Embracing artifact sharing will help to strengthen and mature the field.

Challenges and Opportunities of Artifact Sharing in the Security Field

Benefits of Artifact Sharing Beyond the Research Community





B. Hermann, "What Has Artifact Evaluation Ever Done for Us?," in IEEE Security & Privacy, vol. 20, no. 5, pp. 96-99, Sept.-Oct. 2022, https://doi.org/10.1109/MSEC.2022.3184234

Prior pessimism about reuse in software engineering research may have been a result of using the wrong methods to measure the wrong things.

BY MARIA TERESA BALDASSARRE, NEIL ERNST, BEN HERMANN, TIM MENZIES, AND RAHUL YEDIDA

(Re)Use of Research Results (Is Rampant)

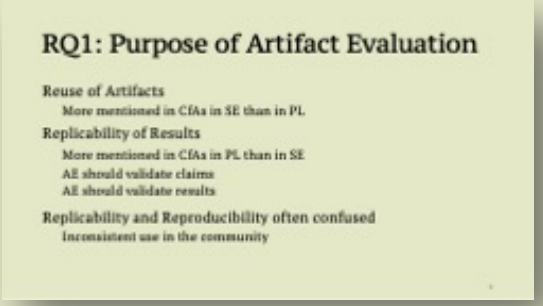
key insights

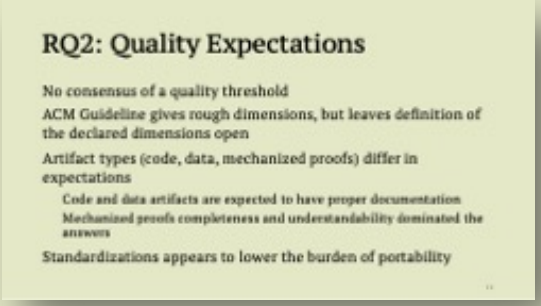
Maria Teresa Baldassarre, Neil Ernst, Ben Hermann, Tim Menzies, and Rahul Yedida. 2023. (Re)Use of Research Results (Is Rampant). Commun. ACM 66, 2 (February 2023), 75–81. https://doi.org/10.1145/3554976

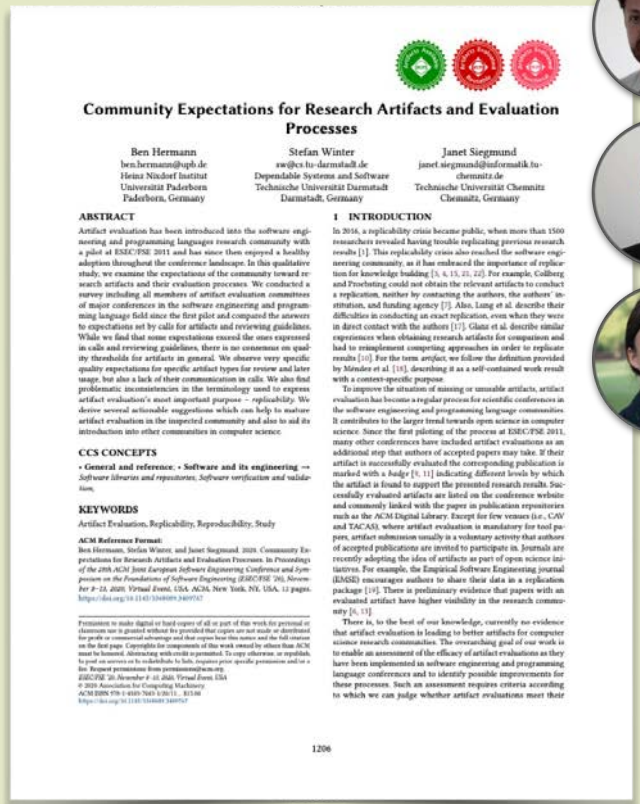
# Community Expectations for Research Artifacts and Evaluation Processes

Ben Hermann, Stefan Winter, Janet Siegmund





Community Expectations for Research Artifacts and Evaluation Processes

Ben Hermann, ben.hermann@upb.de, Heinz Nixdorf Institut, Universität Paderborn, Paderborn, Germany
Stefan Winter, sw@cs.tu-darmstadt.de, Dependable Systems and Software, Technische Universität Darmstadt, Darmstadt, Germany
Janet Siegmund, janet.siegmund@informatik.tu-chemnitz.de, Technische Universität Chemnitz, Chemnitz, Germany

ABSTRACT

Artifact evaluation has been introduced into the software engineering and programming languages research community with a pilot at ESEC/FSE 2011 and has since then enjoyed a healthy adoption throughout the conference landscape. In this qualitative study, we examine the expectations of the community toward research artifacts and their evaluation processes. We conducted a survey including all members of artifact evaluation committees of major conferences in the software engineering and programming language field since the first pilot and compared the answers to expectations set by calls for artifacts and reviewing guidelines. While we find that some expectations exceed the ones expressed in calls and reviewing guidelines, there is no consensus on quality thresholds for artifacts in general. We observe very specific quality expectations for specific artifact types for review and later usage, but also a lack of their communication in calls. We also find problematic inconsistencies in the terminology used to express artifact evaluation's most important purpose – replicability. We derive several actionable suggestions which can help to mature artifact evaluation in the impacted community and also to aid its introduction into other communities in computer science.

CCS CONCEPTS

• General and reference. • Software and its engineering → Software libraries and repositories, Software verification and validation.

KEYWORDS

Artifact Evaluation, Replicability, Reproducibility, Study

ACM Reference Format:
Ben Hermann, Stefan Winter, and Janet Siegmund. 2020. Community Expectations for Research Artifacts and Evaluation Processes. In Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '20), November 8–13, 2020, Virtual Event, USA. ACM, New York, NY, USA. 12 pages. https://doi.org/10.1145/3368089.3409767

1 INTRODUCTION

In 2016, a replicability crisis became public, when more than 1500 researchers revealed having trouble replicating previous research results [1]. This replicability crisis also reached the software engineering community, as it has embraced the importance of replication for knowledge building [3, 4, 13, 21, 22]. For example, Collberg and Proebsting could not obtain the relevant artifacts to conduct a replication, neither by contacting the authors, the authors' institution, and funding agency [7]. Also, Lung et al. describe their difficulties in conducting an exact replication, even when they were in direct contact with the authors [17]. Glatz et al. describe similar experiences when obtaining research artifacts for comparison and had to reimplement competing approaches in order to replicate results [10]. For the term *artifact*, we follow the definition provided by Méndez et al. [18], describing it as a self-contained work result with a context-specific purpose.

To improve the situation of missing or unusable artifacts, artifact evaluation has become a regular process for scientific conferences in the software engineering and programming language communities. It contributes to the larger trend towards open science in computer science. Since the first piloting of the process at ESEC/FSE 2011, many other conferences have included artifact evaluations as an additional step that authors of accepted papers may take. If their artifact is successfully evaluated the corresponding publication is marked with a *badge* [9, 11] indicating different levels by which the artifact is found to support the presented research results. Successfully evaluated artifacts are listed on the conference website and commonly linked with the paper in publication repositories such as the ACM Digital Library. Except for few venues (i.e., CAV and TACAS), where artifact evaluation is mandatory for tool papers, artifact submission usually is a voluntary activity that authors of accepted publications are invited to participate in. Journals are recently adopting the idea of artifacts as part of open science initiatives. For example, the Empirical Software Engineering journal (EMSE) encourages authors to share their data in a replication package [19]. There is preliminary evidence that papers with an evaluated artifact have higher visibility in the research community [8, 13].

There is, to the best of our knowledge, currently no evidence that artifact evaluation is leading to better artifacts for computer science research communities. The overarching goal of our work is to enable an assessment of the efficacy of artifact evaluations as they have been implemented in software engineering and programming language conferences and to identify possible improvements for these processes. Such an assessment requires criteria according to which we can judge whether artifact evaluations meet their

Permission to make digital or hard copies of all or part 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. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ESEC/FSE '20, November 8–13, 2020, Virtual Event, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7043-1/20/11…$15.00
https://doi.org/10.1145/3368089.3409767

1206

**Paper:** https://doi.org/10.1145/3368089.3409767
**Artifact:** https://doi.org/10.5281/zenodo.3951724