FOSD `24

Understanding the current state of replication packages and finding ways to improve their quality and availability

Florian Poreba

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

VERSION: 25.03.2024

Personal Motivation: Replication Study

- **Master thesis:** replication study
- **Negative results**: burden of proof
- **Considerable work**:
  - Tracking all deviations
  - Lack of guidelines
    - Replication package structure
    - Close vs. conceptual replications
    - Hard to find/apply
- **Discouraging context**
  - Replication crisis [1,3]: **Not alone!**
- **Motivation: How could I contribute to fixing this?**

Terminology

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

12

Source: https://benhermann.eu/talks/fse20-expectations-se23.pdf

Artifacts and Packages

- Today: broader term of “artifacts”
  - e.g., Artifacts evaluation, ACM badges
  - Any (digital) sumplementary material: (RAW) data, scripts, documentation, etc. [5]
- Contrast: basili 1999 [6]→ more than just supplementary material
  - Project website (“living document”)
  - Facilitate and track replication
  - Document changes
  - Link related studies

ACM Badges

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

ABSTRACT

Most software-engineering research involves the development of a prototype, a proof of concept, or a measurement apparatus. Together with the data collected in the research process, they are collectively referred to as research artifacts and are subject to arti-

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

RESEARCH-ARTICLE OPEN ACCESS

A retrospective study of one decade of artifact evaluations

Authors: Stefan Winter, Christopher S. Timperley, Ben Hermann, Jürgen Cito, Jonathan Bell, Michael Hilton, Dirk Beyer Authors Info & Claims

ESEC/FSE 2022: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering • November 2022 • Pages 145–156 • https://doi.org/10.1145/3540250.3549172

Published: 09 November 2022 Publication History

Check for updates

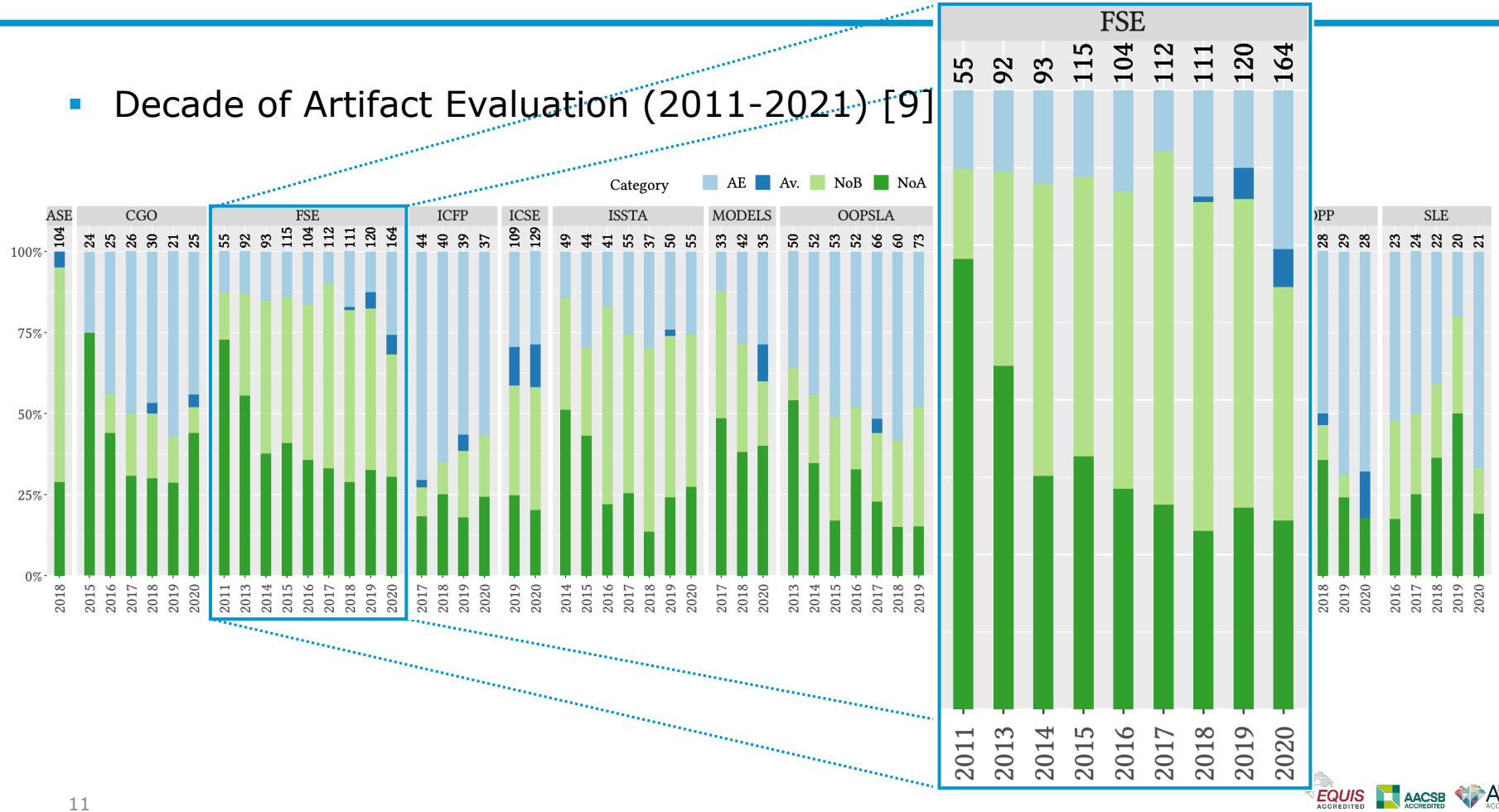
Related Artifact: Reproduction Package (Docker container) for the FSE 2022 Article `A Retrospective Study of one Decade of Artifact Evaluations` • November 2022 • data • https://doi.org/10.5281/zenodo.7082407

1 614

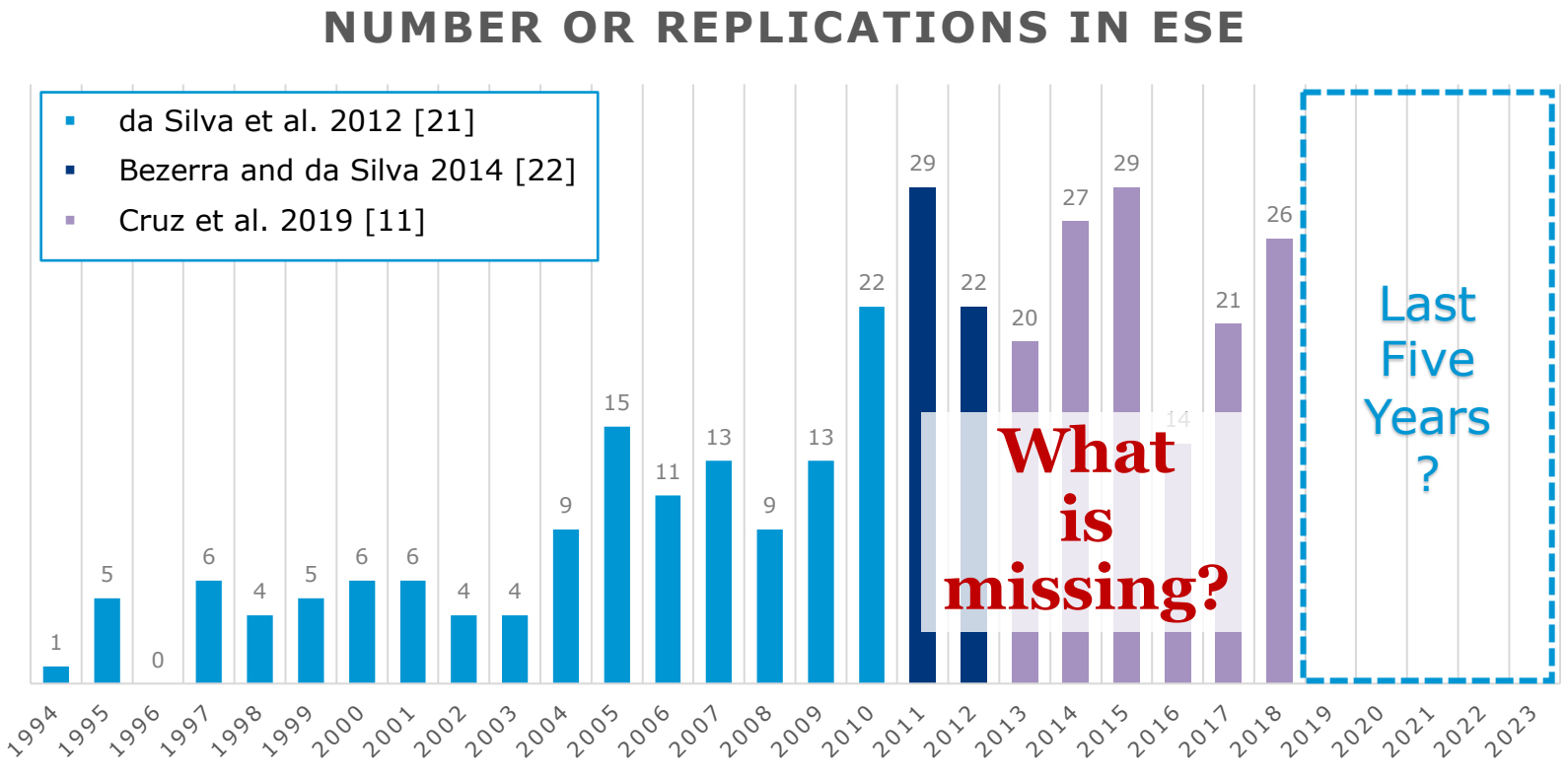
eReader

Current State of Artifacts

- Decade of Artifact Evaluation (2011-2021) [9]



Current State of Replications



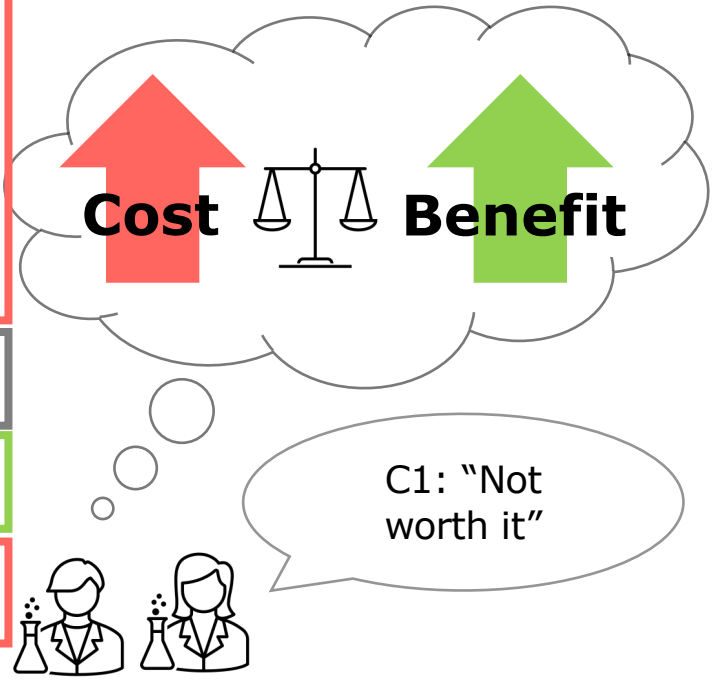
Challenges in Artifact Creation

- Understanding and improving artifact sharing in software engineering research [14]
- 2014-2018: Research Papers at four SE Venues
- Survey of authors
- Three Perspectives
  - Author
  - Reviewer
  - (Re-)User

- C1: Not worth it
- C2: Portability
- C3: Maintenance
- C4: Tacit knowledge
- C5: Artifact does not fit purpose
- C6: Lack of standards and guidelines
- C7: Hosting
- C8: Double-blind review
- C9: External constraints
- C10: Lack of reviewer incentives
- C11: Technical obstacles
- C12: Limited communication

Recommendations

- R1: Describe the Contents, Structure, and Purpose of Artifacts
- R2: Create a Self-Contained Artifact
- R3: Establish a Plan for Creating and Sharing Artifacts
- R4: Obtain and Use a Clear Rubric to Evaluate Artifacts
- R5: Align Author and Reviewer Expectations
- R6: Reduce the Opportunity Cost of Reviewing
- R7: Recognise and Fund Artifact-Related Activities
- R8: Establish a Long-Term Strategy for Artifact Sharing and Evaluation



Past and Current Efforts (non exhaustive)

- Artifact Evaluation (Tracks)
- ACM Badges [5]
- Graph of Re(Use) [15]
- Tool Support
  - TIRA.io [16]
  - SureSoft [10]
  - eXemplar and SEDL (Inactive) [17]
  - CÆSAR (Proof of Concept) [18]
  - Reproducible and Reusable Artifacts (FOSD 2023 – Alina Mailach)
- Zenodo & Figshare Archival (incl. Double-Blind Tutorial [19])

Research Area - Step I: Identifying Status Quo

- SLR Update: How did the number of replications change in the last 5 years?
- SLR: What does the typical replication package look like?
  - Structure?
  - Content?
  - Tools?
  - Standards?
- In what ways are current replication packages still lacking?
  - Challenges and recommendations of Timpreley et al. (see Slides 9&10)
  - Are there other qualities which differentiate „good“ and „bad“ packages?
  - Do the replication packages of replications look different?
    - Gut Feeling: “I wish the replication packages I used looked like this, during my replication”

Research Area - Step II: Support Researchers

- Based on Challenges (C1-C12) and Recommendations (R1-R8) by Timperley et al. [14]
- How can we support researchers?
  - Tools
  - Processes
  - Learning Ressources
  - Communities
  - Incentives / Recognition
- Inspiration from Software Development
  - DevOps
  - Test-Driven Software Development
  - Open Source Maintenance Life Cycles

Questions for the FOSD ‘24 Forum

- Have you noticed any blind spots from my intial investigation?
- What are your experiences on
  - Challenges replicating other work
  - Challenges sharing artifacts
  - Artifact availability and quality
- What are your thoughts on supporting
  - Artifact creation/sharing/maintenance
  - Creating, documenting, maintaining families of experiments
- What are your thoughts on the cost/benefits of artifact sharing?

References I

- [1] John PA Ioannidis. “Why most published research findings are false”. In: PLoS medicine 2.8 (2005), e124.
- [2] Open Science Collaboration. “Estimating the reproducibility of psychological science”. In: Science 349.6251 (2015), aac4716. doi: 10.1126/science.aac4716.
- [3] Monya Baker. “1,500 scientists lift the lid on reproducibility”. In: Nature 533.7604 (2016). doi: 10.1038/533452a.
- [4] Lorena A Barba. “Terminologies for reproducible research”. In: CoRR (2018). doi: 10.48550/arXiv.1802.03311. arXiv: 1802.03311 [cs.DL].
- [5] Association for Computing Machinery. Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-and-badging-current. [Accessed 24-03-2024].
- [6] Victor R. Basili, Forrest Shull, and Filippo Lanubile. “Building Knowledge through Families of Experiments”. In: IEEE Trans. Software Eng. 25.4 (1999), pp. 456–473. doi: 10.1109/32.799939.
- [7] Barbara A. Kitchenham. “The role of replications in empirical software engineering - a word of warning”. In: Empir. Softw. Eng. 13.2 (2008), pp. 219–221. doi: 10.1007/s10664-008-9061-0.
- [8] Martin J. Shepperd, Nemitari Ajienka, and Steve Counsell. “The role and value of replication in empirical software engineering results”. In: Inf. Softw. Technol. 99 (2018), pp. 120–132. doi: 10.1016/j.infsof.2018.01.006.
- [9] Stefan Winter et al. “A retrospective study of one decade of artifact evaluations”. In: ESEC/SIGSOFT FSE. ACM, 2022, pp. 145–156. doi: 10.1145/3540250.3549172.
- [10] Christopher Blech et al. “SURESOFT: Towards Sustainable Research Software”. (2022). doi: 10.24355/dbbs.084-202210121528-0.
- [11] Margarita Cruz et al. “Replication of studies in empirical software engineering: A systematic mapping study, from 2013 to 2018”. In: IEEE Access 8 (2019), pp. 26773–26791. doi: 10.1109/ACCESS.2019.2952191.

References II

- [12] Ben Hermann, Stefan Winter, and Janet Siegmund. "Community expectations for research artifacts and evaluation processes". In: ESEC/SIGSOFT FSE. ACM, 2020, pp. 469–480. doi: 10.1145/3368089.3409767.
- [13] Janet Siegmund, Norbert Siegmund, and Sven Apel. "Views on Internal and External Validity in Empirical Software Engineering". In: ICSE (1). IEEE Computer Society, 2015, pp. 9–19. doi: 10.1109/ICSE.2015.24.
- [14] Christopher S. Timperley et al. "Understanding and improving artifact sharing in software engineering research". In: Empirical Software Engineering 26.4 (2021), p. 67. doi: 10.1007/s10664-021-09973-5.
- [15] Maria Teresa Baldassarre et al. "(Re)Use of Research Results (Is Rampant)". In: Commun. ACM 66.2 (Jan. 2023), pp. 75–81. issn: 0001-0782. doi: 10.1145/3554976.
- [16] Maik Fröbe et al. "Continuous Integration for Reproducible Shared Tasks with TIRA.io". In: ECIR (3). Vol. 13982. Lecture Notes in Computer Science. Springer, 2023, pp. 236–241. doi: 10.1007/978-3-031-28241-6_20.
- [17] José Antonio Parejo Maestre et al. "EXEMPLAR: An Experimental Information Repository for Software Engineering Research". (2014).
- [18] Margarita Cruz et al. "A model-based approach for specifying changes in replications of empirical studies in computer Science". In: Computing 105.6 (2023), pp. 1189–1213. doi: 10.1007/s00607-022-01133-x.
- [19] Daniel Graziotin. dgraziotin/disclose-data-dbr-first-then-opendata: v1.0.1. Version v1.0.1. Jan. 2024. doi: 10.5281/zenodo.10532090.
- [20] Omar S. Gómez, Natalia Juristo Juzgado, and Sira Vegas. "Understanding replication of experiments in software engineering: A classification". In: Inf. Softw. Technol. 56.8 (2014), pp. 1033–1048. doi: 10.1016/j.infsof.2014.04.004.
- [21] Da Silva, F. Q., Suassuna, M., França, A. C. C., Grubb, A. M., Gouveia, T. B., Monteiro, C. V., & dos Santos, I. E. (2014). Replication of empirical studies in software engineering research: a systematic mapping study. Empirical Software Engineering, 19, 501-557. doi: 10.1007/s10664-012-9227-7.
- [22] R. M. M. Bezerra, F. Q. B. da Silva, A. M. Santana, C. V. C. Magalhaes and R. E. S. Santos, "Replication of Empirical Studies in Software Engineering: An Update of a Systematic Mapping Study," 2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), Beijing, China, 2015, pp. 1-4, doi: 10.1109/ESEM.2015.7321213.

Thank you for your attention!

VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Department of Information Systems and Operations Management
Institute for Complex Networks
Welthandelsplatz 1, 1020 Vienna, Austria

Florian Poreba, MSc (WU)

T +43-1-31336-5017
florian.poreba@wu.ac.at
https://complex.wu.ac.at/nm/en:poreba