

An Overview of Competency Areas in Artifact Evaluation

Martin Armbruster ¹ and Jan Bernoth ²

Abstract: For assessing the quality of (FAIR) research artifacts, artifact evaluations as a form of peer reviews have been established. However, they raise questions about the required competencies. Thus, this short paper provides an overview of six artifact evaluation processes and their goals, which exhibit a multidimensional influence on the required competencies. Based on these factors, initial competency areas can be identified.


Keywords: Artifact Evaluation, Competencies, Research Software Engineering, Research Data Management, National Research Data Infrastructure for and with Computer Science

1 Introduction

The discourse surrounding the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [Wi16] has centered on assisting researchers in creating FAIR-compliant artifacts to support open and trustworthy research. Central discussion points are, especially, how to enrich research output with metadata and how to enhance reusability. One aspect which is mostly left out is how to implement quality assessment mechanisms within the research data management lifecycle [BS22], which can also enable re-use.

In the following, research output will be referred to as artifacts which are either “used as part of the study or generated by the experiment itself” (Association for Computing Machinery (ACM) [As20]). Quality assessment of research artifacts represents a multifaceted challenge that requires both automated mechanisms (e.g., algorithmic analysis or reproducibility checks) and expert-driven evaluation processes (i.e., human judgment based on domain knowledge and scholarly standards). Quality assessment is not a novel concept for researchers. Indeed, numerous processes within the academic ecosystem incorporate review mechanisms, predominantly conducted through peer evaluations. However, this raises important questions. To which extent do competencies required for effective *artifact evaluations* (AE) overlap with those already embedded in existing evaluation processes? Is there a need for a specialized set of competencies specifically tailored to AE processes? If so, which competencies are contained in such a set? In the end, a question is also which technical support is necessary or desirable to enhance the competencies of reviewers so that they can focus rather on the review and less on the technicalities of the artifacts under evaluation. To open the discussion on these questions, this short paper provides a non-exhaustive overview of current AE

¹ Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131 Karlsruhe, Germany,
martin.armbruster@kit.edu,  <https://orcid.org/0000-0002-2554-4501>

² University of Potsdam, Department of Computer Science, An der Bahn 2, 14476 Potsdam, Germany,
jan.bernoth@uni-potsdam.de,  <https://orcid.org/0000-0002-4127-0053>

processes and their goals and explores related, potential competency areas necessary for conducting thorough artifact assessments, targeted towards AE organizers.

2 Types of Artifact Evaluation Processes

AE takes various forms across the Research Software Engineering (RSE) landscape, serving different purposes while sharing the common goal of ensuring quality, reproducibility, and sustainability of research outputs.

1. *Peer Review within a Team:* One of the most established forms of AE occurs within RSE teams through code review processes. These internal evaluations help maintain code quality and adherence to best practices before software is released.
2. *AE Tracks in Conferences:* At a more formal level, many academic conferences include dedicated AE tracks. For example, the SuperComputing 2025 [SC] and International Conference on Software Engineering (ICSE) 2025 [IC] have established rigorous evaluation processes that assess the reproducibility, functionality, and documentation of software and research artifacts accompanying research papers. After successful evaluation, authors are rewarded with badges (e.g., ACM badges [As20]), highlighting the performed AE.
3. *Publication:* Beyond conference evaluations, dedicated publication venues for software have emerged. The Journal of Open Source Software³ provides peer review and publication specifically for research software, focusing on code quality, documentation, and community standards. Similarly, the Dagstuhl Artifacts Series⁴ offers a platform for publishing and preserving digital artifacts related to computer science research.
4. *Archiving:* Software preservation initiatives such as the Software Heritage Foundation⁵ also incorporate evaluation processes in their moderation workflows [Di20], ensuring that archived software meets certain quality and documentation standards before being permanently preserved.
5. *Recognition Systems:* Recognition systems have been developed to incentivize high-quality research software. E.g., the Helmholtz Software Award⁶ recognizes outstanding software development in scientific contexts and promoting open-source practices.
6. *Funding:* Consequentially, funding agencies increasingly incorporate software evaluation into their decision-making processes [Ka23]. Whether from public agencies, philanthropic organizations, or commercial entities, funding for research software development often depends on rigorous evaluation of software quality, sustainability plans, and potential impact.

³ <https://joss.theoj.org/>

⁴ <https://www.dagstuhl.de/publishing/series/details/darts>

⁵ <https://archive.softwareheritage.org/>

⁶ <https://os.helmholtz.de/en/open-research-software/helmholtz-software-award/>

These diverse forms of AEs result in different process characteristics and evaluation goals, as described in the next section 3.

3 Characteristics and Goals of Artifact Evaluations

Each AE process encompasses *various characteristics*, which are influencing factors on the competencies needed for AE. In addition, certain characteristics are also connected to the goals of the AE. For example, the *types of the artifacts* affect the concrete review step and goal by how the artifacts are reviewed. If the artifacts only consist of data, the evaluation could cover the completeness and consistency of the data. On the other hand, if the artifacts include or solely contain software, the evaluation is (partly) a code review.

Another important characteristic is the *requirement and expectation on the time* spent for the evaluation. While, for instance, sufficient time should be allocated for peer reviews within a team and publications to thoroughly review the artifacts, AE tracks in conferences impose time limits on reviews [IC; SC] to streamline the AE process.

Furthermore, one characteristic is the *form of communication*, affected by two aspects. First, usually text-based, the *granularity of the reviews* differs. In peer reviews within a team and publications, reviewers employ fine-grained and line-based comments for feedback. In contrast, more coarse-grained, summarizing reviews are found in AE tracks in conferences, archiving, recognition systems, or funding. The second aspect deals with the *relationship between authors and reviewers*. In peer reviews within a team, they likely know each other. Contrary, authors and reviewers do not likely know each other in the case of a publication. At last, AE tracks in conferences usually enforce a single-blind review so that authors do not know who the reviewers are. In summary, both aspects determine how authors and reviewers should communicate.

As indicated before, each type of AE serves different purposes and goals. In the context of AE tracks in conferences, Hermann; Winter; Siegmund [HWS20] conducted a survey among AE committee members on their perceived purposes for the AE. They identified the following two groups and included purposes / goals:

- Fostering properties of artifacts: Reproducibility, reusability, comparability, repeatability, replicability, usability, and availability
- Checking properties of the artifact: Validating claims, validating results, validating reusability, validating reproducibility, validating existence, validating replicability, and validating usability

These results suggest that reviewers in AE tracks consider multiple goals [HWS20]. Depending on the artifact types and badges, a reviewer employs a particular combination of these goals for every evaluated artifact, while the main focus lies on validating claims and validating results [HWS20]. Both groups, *fostering properties of artifacts* and *checking*

properties of the artifact, are also applicable to the other forms of AEs with additional goals and shifted focuses. For example, in peer reviews within a team, publication, and archiving, the focus is on checking and validating that corresponding code quality and documentation standards are met. Additionally, code quality can be broken down into further characteristics (e.g., security or maintainability according to ISO/IEC 25010:2023 [IS23]). At the same time, peer reviews within a team, publications, and archiving foster the properties of artifacts, as reviewers can suggest improvements.

4 Discussion and Outlook

As outlined, there are several factors stemming from the AE processes and goals, which exhibit a multidimensional influence on the competencies for AEs and which need to be considered. Thus, based on these factors, the following initial, broad competency areas for AEs can be derived (inspired by Wurzel Gonçalves et al. [Wu23]): subject-specific (e.g., for validating claims or results), research-related (e.g., for validating soundness), technical (e.g., for assessing software quality), social (e.g., for writing reviews or comments as a form of communication), and personal (e.g., for time management). Nevertheless, the extent to which single competencies are utilized during an AE varies. For example, while AE tracks in conferences require more subject-specific and research-related and less technical competencies, recognition and software publication systems tend to demand more technical competencies. In general, underlying infrastructures should automate routine checks, such as format validation or metadata completeness, to reduce the workload of experts in AEs.

Given the initial competency areas, an additional question raises: do all reviewers need to master all competencies? If not, this can open possibilities for alternative AE processes and interdisciplinary collaborations. As one example, there could be one reviewer for the subject and one reviewer for technical aspects. As a second example, there could be a two-step process. At first, artifacts are reviewed in an AE track, focusing on the subject. Next, a separate publication, concentrating on the technical aspects, takes place.

To conclude, the discussions above imply that there is an overlap in the competencies of AEs and other peer evaluations. However, additional competencies and areas, in particular, the technical one, are required. In future work, these insights can become part of a software literacy framework, utilizing results from existing literature (e.g., on competencies for code reviews [Wu23] or paper reviews [Kö20]). Additionally, recommendations for organizers of AE processes can be created to incorporate the identified factors.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 52/1 – 501930651. Generative AI assisted in the formulation of section 1 and section 2.

Bibliography

- [As20] Association for Computing Machinery: Artifact Review and Badging - Current, 2020, <https://www.acm.org/publications/policies/artifact-review-and-badging-current>, visited on: 05/15/2025.
- [BS22] Biernacka, K.; Schulz, S.: Forschungsdatenmanagement in der Informatik. Logos Verlag, Berlin, 2022.
- [Di20] Di Cosmo, R. et al.: Curated Archiving of Research Software Artifacts: Lessons Learned from the French Open Archive (HAL). *International Journal of Digital Curation* 15 (1), p. 16, 2020.
- [HWS20] Hermann, B.; Winter, S.; Siegmund, J.: 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. ESEC/FSE 2020*, Association for Computing Machinery, Virtual Event, USA, pp. 469–480, 2020, <https://doi.org/10.1145/3368089.3409767>.
- [IC] ICSE 2025: Artifact Evaluation, <https://conf.researchr.org/track/icse-2025/icse-2025-artifact-evaluation>, visited on: 05/15/2025.
- [IS23] ISO/IEC 25010:2023(E): Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model, Standard, Geneva, CH: International Organization for Standardization, 2023.
- [Ka23] Katz, D. S.: Evaluating research software proposals, 2023, <https://danielskatzblog.wordpress.com/2023/02/27/evaluating-research-software-proposals/>, visited on: 05/15/2025.
- [Kö20] Köhler, T. et al.: Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: Introducing a competency framework for peer review. *Industrial and Organizational Psychology* 13 (1), pp. 1–27, 2020.
- [SC] SC25: AD/AE Process & Badges, <https://sc25.supercomputing.org/program/papers/reproducibility-appendices-badges/>, visited on: 05/15/2025.
- [Wi16] Wilkinson, M. D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (1), 2016, <https://doi.org/10.1038/sdata.2016.18>.
- [Wu23] Wurzel Gonçalves, P. et al.: Competencies for Code Review. *Proc. ACM Hum.-Comput. Interact.* 7 (CSCW1), 2023, <https://doi.org/10.1145/3579471>.