FSE 2021 Artifact Evaluation Track: Submission and Reviewing Guidelines

Silvia Abrahão and Daniel Mendez, amended for FSE 2021 by Panos Louridas

March 23, 2021

Scope and Objectives

This document outlines, briefly, the submission and reviewing process for the Artifact Evaluation track of FSE 2021. It aims at providing authors and reviewers with pragmatic insights into the process and expected criteria to merit awarding the respective badges.

General Remarks on the Artifact Evaluation Track and Expected Attitude

In principle, the goal of the track is to promote and celebrate open science. We therefore understand the track as one important means to actively engage with the community in order to support them in making their research artifacts publicly available and in fostering replication of research results. The final result of the artifact evaluation is to reward (only) the authors' work that satisfies the criteria listed below with a set of qualifying badges as a form of recognition. Yet, we see the track and the review phase as a unique chance to actively support the research community in open science, so instead of reviewing the artifacts "blindly" according to the evaluation criteria towards the end of the review phase and submitting a review with a go/no-go decision, we encourage all reviewers to make use of the rebuttal phase in order to actively support the authors in improving their submissions, same as we encourage authors to actively engage with the reviewers and do their best to address their well-intended suggestions as efficiently as possible.

Contents

1 Badges Overview and Eligibility

The artifact evaluation track aims to review, promote, share, and catalog the research artifacts of accepted software engineering papers. Authors of an accepted *research paper* can submit an artifact for the Artifacts Evaluated and Artifacts Available badges. Authors of any prior SE work (published at FSE or elsewhere) are also invited to submit their work for the Results Validated badges. Definitions for the badges are given in the table below, taken from the ACM Artifact Review and Badging Version 1.1. The top two artifacts selected by the Programme Committee will be awarded the best artifact awards. All accepted abstracts documenting the artifacts will be published in the FSE 2021 proceedings as a further form of recognition.

Artifacts Evaluated		Artifacts Available	Results Validated	
Functional	Reusable		Results Reproduced	Results Replicated
The artifacts associated with the research are found to be documented, consistent, complete, exercisable, and include appropriate evidence of verification and validation.	Functional + the artifacts associated with the paper are of a quality that significantly exceeds minimal functionality. They are very carefully documented and well-structured to the extent that reuse and repurposing is facilitated. In particular, norms and standards of the research community for artifacts of this type are strictly followed.	Author-created artifacts relevant to this paper have been placed on a publicly accessible archival repository. A DOI or link to this repository along with a unique identifier for the object is provided.	The main results of the paper have been obtained in a subsequent study by a person or team other than the original authors, using, in part, artifacts provided by the original authors.	The main results of the paper have been independently obtained in a subsequent study by a person or team other than the original authors, without the use of author-supplied artifacts.

Badges in More Detail

There are three different badges, two of which distinguish between two levels.

Artifacts Evaluated: This badge is applied to papers whose associated artifacts have successfully completed an independent audit. Artifacts need not be made publicly available to be considered for this badge. However, they do need to be made available to reviewers. Two levels are distinguished, only one of which should be applied in any instance:

- Artifacts Evaluated - Functional These artifacts need to be:

    - documented: At minimum, an inventory of artifacts is included, and sufficient description provided to enable the artifacts to be exercised.

    - consistent: The artifacts are relevant to the associated paper, and contribute in some inherent way to the generation of its main results.

    - complete: To the extent possible, all components relevant to the paper in question are included. (Proprietary artifacts need not be included. If they are required to exercise the package then this should be documented, along with instructions on how to obtain them. Proxies for proprietary data should be included so as to demonstrate the analysis.)

    - exercisable: Included scripts and / or software used to generate the results in the associated paper can be successfully executed, and included data can be accessed and appropriately manipulated.

- Artifacts Evaluated - Reusable The artifacts meet the requirements for the Artifacts Evaluated - Functional level and in addition they are of a quality that significantly exceeds the requirements set for the first level. *Authors are strongly encouraged to target their artifact submissions for Artifacts Evaluated - Reusable*, as the purpose of artifact badges is, among other things, to facilitate reuse and repurposing, which may not be achieved at the Artifacts Evaluated - Functional level.

Artifacts Available: This badge is applied to papers in which associated artifacts have been made permanently available for retrieval.

- We consider temporary drives (e. g., Dropbox, Google Drive) to be non-persistent, same as individual/institutional websites of the submitting authors, as these are prone to changes.

- We do not mandate the use of specific repositories. Although not limited to, we strongly recommend relying on services like Zenodo to archiving repositories / repository releases (e. g., from GitHub) as these services are persistent and they also offer the possibility to assign a DOI. In principle, however, publisher repositories (e. g., ACM Digital Library) and open commercial repositories (e. g., figshare) are acceptable as well as long as they offer a declared plan to enable permanent accessibility.

- Artifacts do not need to have been formally evaluated in order for an article to receive this badge. In addition, they need not be complete in the sense described above. They simply need to be rele-

4

vant to the study and add value beyond the text in the article. Such artifacts could be something as simple as the data from which the figures are drawn, or as complex as a complete software system under study.

Results Validated: This badge is applied to papers in which the main results of the paper have been successfully obtained by a person or team other than the author. Two levels are distinguished, only one of which should be applied in any instance:

- Results Replicated The results were validated by a person or team other than the original authors of the work, with, at least in part, artifacts provided by the original authors.

- Results Reproduced As in Results Replicated, but without any artifacts provided by the original authors.

Examples:

- If Asha published a paper with artifacts in 2019, and Tim published a replication in 2020 using the artifacts, then Asha can now apply for the Results Replicated badge on the 2019 paper.

- If Cameron published a paper in 2018 with no artifacts, and Miles published a paper with artifacts in 2020 that independently obtained the main result, then Cameron can apply for the Results Reproduced badge on the 2018 paper.

If the artifact is accepted as Results Validated:

- Authors will be invited to give lightning talks on this work at the ROSE session at FSE 2021. The ROSE (Recognizing and Rewarding Open Science in Software Engineering) festival is a world-wide salute to replication and reproducibility in software engineering. Our aim is to create a venue where researchers can receive public credit for facilitating and participating in open science in software engineering (specifically, in creating replicated and reproduced results).

- We will work with the IEEE Xplore and ACM Portal administrator to add badges to the electronic versions of the paper related to the artifact.

2 Submission Process

Submission Overview

In principle, authors are expected to submit through HotCRP their artifact documentation. This documentation distinguishes two basic types of information—captured in one central research abstract (two pages max)—depending on the intented badge.

Artifacts Evaluated: The emphasis lies on providing documentation on the research artifact previously prepared and archived. Here, the authors need to write and submit documentation explaining how to obtain the artifact package, how to unpack the artifact, how to get started, and how to use the artifact in more detail. The submission must only describe the technicalities of the artifact and uses of the artifact that are not already described in the paper.

Artifacts Available: The authors must give the location of the artifact on a publicly acces-

sible archival repository, along with a DOI or a link to the repository. This means that the HotCRP submission should include the research abstract only providing links to the repositories where the artifact is permanently stored and available. Submitting artifacts themselves through HotCRP without making them publicly accessible (through a repository or an archival service) will not be sufficient.

Results Validated: The emphasis here lies on providing information about how their already published research has been replicated or reproduced as well as links to further material (e. g., the papers and artifacts in question). We encourage submissions for those badges by the replicating authors nominating the original authors.

! → If the authors are not aiming for the Artifacts Available badge, the artifacts do not necessarily have to be publicly accessible for the review process. However, *the authors should clearly explain why the artifact are not publicly available*, for example, because of privacy concerns, law, or NDAs in place. In this very case, the authors are asked to provide either a private link / password-protected link to a repository or they may submit the artifact directly through HotCRP (in a zip file) and it should become clear which steps are necessary for authors who would like to reuse the artifact.

Types of Research Artifacts

There are two options depending on the nature of the artifacts: Instalation Package or Simple Package. In general, an Instalation Package is related to software artifacts or, for instance, scripts, while a Simple Package may be related to qualitative studies (e. g., interview transcripts or coding schemas).

In both cases, it is expected that the basic set-up of the artifact (including configurations and installations) take less than 30 minutes. Otherwise, the artifact is unlikely to be explicitly endorsed by Program Committee members because they will simply will not have enough time to deal with it.

Instalation Package If the artifact consists of a tool or software system, then the authors need to prepare an Installation Package so that the tool can be installed and run in the evaluator's environment. That is to say, please make sure to provide enough associated instructions, code, and data such that any Software Engineering person with a reasonable knowledge of scripting, build tools, etc., could install, build, and run the code. If the artifact contains or requires the use of a special tool or any other non-trivial piece of software, the authors must provide a VirtualBox VM image or a Docker container image with a working environment containing the artifact and all the necessary tools. We expect that the artifacts have been vetted on a clean machine before submission.

Simple Package If the artifact contains documents that can be used with a simple text editor, a PDF viewer, or some other common tool (e. g., a spreadsheet program in its basic configuration) the authors can just save all documents in a single package file (zip or tar.gz).

Artifact Documentation

Regardless of the badge, authors must provide documentation explaining how to obtain the artifact package, how to unpack the artifact, how

6

to get started, and how to use the artifacts in more detail. The artifact itself must only describe the technicalities of the artifact and uses of the artifact that are not already described in the paper; nevertheless, the artifact and its documentation should be self-contained. The submission should contain (and / or link to) the documents listed below. The documents should be in plain text, MarkDown, or PDF format, indicated by the file extension. The name of each file should be in capital letters.

- A README main file describing what the artifact does and where it can be obtained (with hidden links and access password if necessary). Also, there should be a clear description how to repeat, replicate, or reproduce the results presented in the paper. Artifacts that focus on data should, in principle, cover aspects relevant to understand the context, data provenance, ethical and legal statements (as long as relevant), and storage requirements. Artifacts that focus on software should, in principle, cover aspects relevant to how to install and use it (and be accompanied by a small example).

- A REQUIREMENTS file for artifacts that focus on software. This file should, in principle, cover aspects of hardware environment requirements (e. g., performance, storage or non-commodity peripherals) and software environments (e. g., Docker, VM, and operating system). If relevant, any additional file with version-specific dependencies information (e. g., requirements.txt for Python-only environments, Cargo.toml for Rust, etc.), should be included according to the norms of the specific language and platform. Any deviation from standard environments needs to be reasonably justified.

- A STATUS file stating what kind of badge(s) the authors are applying for as well as the reasons why the authors believe that the artifact deserves that badge(s).

- A LICENSE file describing the distribution rights. Note that for the Artifacts Available badge the artifact needs to be under some form of open source license.

- An INSTALL file with installation instructions. These instructions should include notes illustrating a very basic usage example or a method to test the installation. This could be, for instance, on what output to expect that confirms that the code is installed and working; and the code is doing something interesting and useful.

- A copy of the accepted paper in PDF format.

3 Review Process

This section's intended audience is the Program Committee and, thus, addresses the Program Committee members of the Artifact Evaluation track (and is written accordingly), but it is available to authors as well to facilitate transparency.

The tasks of the reviewers of research artifacts involve three phases:

1. Bidding Phase (May 29–June 4, 2021)

2. Initial Review and Rebuttal Phase (June 5–June 19, 2021)

3. In-depth Review Phase (June 19–July 5, 2021)

Bidding Phase (May 29–June 4)

→ Section 2

Authors who are planning to submit a research artifact are requested to register their artifacts by May 28, 2021 using HotCRP. The submission includes a research abstract with all relevant information and / or links to the repositories containing the information (such as the artifact itself). In the exceptional cases described above, the artifact itself may also be submitted as a zip through HotCRP. For more details, please see the submission process described above in Section 2. Immediately after the submission deadline, we will invite the reviewers to submit their bids in the HotCRP tool.

📅 The bidding deadline is June 4, 2021.

Reviewers should consider their conflicts of interest, research topics, and experiences with specific tools and technologies (if applicable) when placing their bids.

Initial Review and Rebuttal Phase (June 5–June 19)

Authors will submit their artifacts by June 4, 2021. We will then assign artifacts to reviewers as soon as possible.

Before the actual In-depth Review Phase (where no interaction with the authors will take place anymore), reviewers will be asked to check the integrity of the research artifacts and to look for possible setup problems or other smaller technical issues that may prevent the artifact from being properly evaluated (e. g., corrupted or missing files, provided VMs won't start, immediate crashes on the simplest example). During this phase, Program Committee members may contact the authors to request clarifications on the basic installations and start-up procedures or to resolve simple installation problems. Reviewers who wish to communicate with the authors of the artifacts are asked to email the track chairs. In this case we will send the authors and the reviewers a URL to access a chat allowing them to *communicate anonymously* during the rebuttal period. The tool we will use for the communication during the Initial Review and Rebuttal Phase is Etherpad. The orchestration of the communication is done by the Program Committee chairs.

To expedite the review process, we are encouraging the reviewers to try to send all their issues related to installation in one short message, if possible. Given the short review time available, *the authors are expected to respond within a 48-hour period*.

We plan to make any communication between a reviewer and the authors visible to other reviewers assigned to the same artifact to mitigate unnecessary overlaps in effort.

📅 The initial review and rebuttal phase will end on June 19, 2021.

In-depth Review Phase (June 20–July 5)

→ Section 2

After the first quick checks during the initial review and rebuttal phase, possibly leading to the fixing of problems or clarifications during the initial review and rebuttal phase, the actual in-depth review will start. We will use a single-blind review process.

Reviewers review the artifact documentation provided by the authors (e. g., referring to the README file in a repository). Section 2 provides

further details about the expected outline of the research artifacts. Except for exceptional cases, the files comprising the artifact and described in the abstract should already be publicly accessible through a repository. In exceptional cases, however, authors might have submitted the files as a package (e.g. zip) through HotCRP: those cases refer primarily to the cases where authors do not apply for the Artifacts Available badge and where public disclosure of the artifact is not possible, e. g., due to NDAs.

The authors should explain in their submission which badges they are aiming for (STATUS file). The reviewers are then asked to review the artifact for the respective criteria (see Section 4) and decide whether the envisioned badge(s) can be awarded, whether an alternative badge should be awarded (provided the submission meets the criteria), or whether no badge can be awarded at all.

Reviewers are expected to assess if and how the things described in the abstract submission are reflected by the actual artifact in the repository. However, we would like to stress the importance to avoid a black and white decision or searching for small issues that prevent issuing a badge. *The whole point of this track is to promote open science in our research community and help authors willing to share their artifacts in doing this correctly (and efficiently)*.

Reviewers are expected to enter the badge decision on HotCRP together with a short review explaining the badge decision. Please note that we do not expect an in-depth review report, but only a short explanation why or why not a certain badge should be awarded. Furthermore, note that a paper can receive multiple badges.

Artifacts may be awarded one, two, or all three of the Artifacts Evaluated, Artifacts Available, and Results Validated badges. You can therefore in HotCPR all of the scores that apply:

- NO BADGE

- FUNCTIONAL or REUSABLE

- AVAILABLE

- REPRODUCED or REPLICATED

! → Reviewers are asked to submit their reviews as soon as possible and not to submit all their reviews at once at the end of the review phase. We allow discussions between reviewers to take place at any time during the review phase and *all reviews will be made visible to all reviewers of the same artifact as soon as they are submitted* to facilitate effective discussions (and feedback/support by other reviewers) and, again, to mitigate unnecessary overlaps in effort (e. g., to allow reviewers to concentrate on other submissions first).

Finally, it is allowed to involve an external reviewer in cases the reviewer would like to obtain additional feedback or expertise. In that case, it is important to stress the confidentiality of the process to the external reviewer. However, reviewers are expected to also familiarize themselves with the research artifact such that they can assess it fairly. Regardless of the eventual involvement of external reviewers, please note that *the Program Committee members assigned to the artifact are personally responsible for the reviews (with respect to their fairness and accuracy of*

the decision)! Furthermore, we expect the Program Committee members to personally participate in the online discussion.

Nominations: If you want to nominate a research artifact for the best artifact award, please do so by marking it in the review form.

📅 The deadline for submitting reviews is July 5, 2021. Authors will be notified about the decision on July 9, 2021.

Summary of Important Dates

The timeline for the artifact evaluation track is as follows (All dates are 23:59:59 AoE (UTC-12h).

- Friday May 21, 2021: FSE 2021 research paper notification
- Friday May 28, 2021: Artifact pre-submission registration deadline
- Saturday May 29, 2021: Artifact Evaluation bidding start
- Friday June 4, 2021: Artifact Evaluation bidding deadline
- Friday June 4, 2021: Artifact submission deadline
- Saturday June 5, 2021: Start of initial review and rebuttal period
- Saturday June 19, 2021: End of initial review and rebuttal period
- Sunday June 20, 2021: Start of in-depth review phase
- Monday July 5, 2021: Artifact Evaluation review submission deadline
- Friday July 9, 2021: Artifact notification

The Artifact Evaluation notification is only about a month before the conference starts. It is, thus, essential to stick with this schedule!

4 Evaluation Criteria

The subsequent checklist contains a non-exhaustive set of criteria for the evaluation of the artifact submissions for eligibility of the respective badges. We distinguish minimum criteria (which must be met to merit receiving the badge) and optional criteria which we recommend, but do not impose as imperative.

Artifacts Evaluated

Reusable are an extension of Functional. That is, artifacts which qualify for Reusable, are per definition Functional but not necessarily vice-versa. As the scope of the Artifact Evaluation track is to foster reusability of artifacts, authors are expected to submit for the Reusable badge.

Minimum Criteria

- ☑ Artifacts are well documented and offer, at a minimum, an inventory of the contents and sufficient description to enable the artifacts to be exercised.

- ☑ Artifacts are relevant to the associated paper and contribute to the generation of its main results.

10

☑ Artifacts are self-contained and exercisable and include scripts and / or software used to generate the results described in the associated paper, i. e., their integrity allows for a successful execution (if applicable, i. e., software-related) and included data can be accessed and appropriately manipulated.

☑ Artifacts have a proper license available for the artifact, explicitly documented in the LICENSE file.[1]

☑ Instalation Packages have an explicit documentation of the requirements / prerequisites necessary for potential installations or executions of code in the REQUIREMENTS file. This also includes requirements towards operating systems and hardware.

☑ Instalation Packages have an installation script and step-by-step instructions that allow for the automatic installation of necessary tools and environments. When environments or operating systems deviate from the norm (which is essentially always the case as there is no real norm), the package must include as well virtual environments (e. g., Docker container image or VirtualBox VM image). The installation must be executable without problems.[2]

Optional Criteria

☑ Artifacts have an indication of the time needed to run them (e. g., 1 hour, 4 hours, 2 days) and how to run a shorter version (e. g., 10 min.) to check that it is functional.

! → The identification of potential causes for failed installations or executions is not part of the reviewers' tasks.

Artifacts Available

Available artifacts must be made permanently and publicly available, i. e., they are publicly available through a preserved, publicly accessible repository with a stable URL and a DOI.

Minimum Criteria

☑ Artifacts are available for public download from a repository without the need to register.

☑ Artifacts are available for public download from a persistent repository with a stable URL.

☑ Each artifact is associated with a Digital Object Identifier (DOI).

---

[1] The license should indicate the underlying license model (e. g., Creative Commons or MIT) and potential restrictions. The license text should further be self-contained (e. g., by adding the license text as proposed by, for example, CC BY to the LICENCE file). For software, we encourage the use of any open source license. For data, we recommend a Creative Commons license. In any case, the license should allow reuse for scientific and research purposes.

[2] It is the responsibility of submitting authors to provide an installation package that allows to run the artifact in the evaluator's environment. The instructions themselves should be kept to the absolutely required minimum and we recommend relying on virtual environments and automation as much as possible. If the submission includes a simple package with textual files only (e. g., PDFs or spreadsheets), then these documents can be archived in a single package (e. g., zip or tar.gz). The underlying assumption is that if artifacts cannot be installed or exercised without reasonable technical knowledge or without expertise in the research field, then other authors who would make use of that artifact may run into problems as well. In this case, we argue, the badge should not be awarded.

Optional Criteria

- ☑ Artifacts provide explicit documentation on the authors and, ideally, instructions and templates on how to cite when making use of the artifacts. The authors lists are directly accessible from the main description of the artifact or available through a dedicated file (e. g., AUTHORS).

Results Validated

The criteria for the Results Validated badges are primarily assessed based on the submitted research abstracts that outline that (and how) selected artifacts have reached that stage. That is, reviewers are not expected to review the actual reproduction entirely and we expect the abstracts to show that:

- ☑ For Results Replicated, the main results of the paper have been obtained in a subsequent study by a person or team other than the original author, using, in part, the artifacts provided by the author.

- ☑ For Results Reproduced, the main results of the paper have been independently obtained in a subsequent study by a person or team other than the original authors, without the use of author-supplied artifacts.

The main difference between Results Replicated and Results Reproduced lies, therefore, in whether the external replication (partially) needs to rely on artifacts by the authors of the research being replicated or whether the reproduction can be achieved completely independently.

Minimum Criteria

- ☑ The paper reporting on the replication / reproduction has been peer-reviewed.

- ☑ The original paper being reproduced and potentially awarded the badge is publicly available (via a submitted URL directory).

- ☑ The authorship of the reproduced / replicated artifact must not overlap with the reproducing / replicating artifact.

- ☑ The abstract clearly outlines WHAT is being reproduced, WHY it is important, and HOW exactly it has been done. If the replication / reproduction was only partial, then the authors clearly explain what parts could be achieved or which are missing.

- ☑ The submission lays out substantial evidence for replication / reproduction.

- ☑ For Results Replicated only: The abstract clearly shows that the main results of the paper have been obtained without author-supplied artifacts.

Optional Criteria

- ☑ Authors pay due respect to the other work related to the reproduction / replication. That is, the abstract is not necessarily critical towards others in the research community.

☑ Mostly only in case the submitting authors are not the ones of the original work being reproduced / replicated but authors nominating original work: Authors provide a critical reflection upon what aspects made it easier or harder to replicate / reproduce and what are the lessons learned from this work that would enable more replication / reproduction in the future for other kinds of tasks or other kinds of research.

Remarks

To merit the badge Results Replicated or Results Reproduced, it is sufficient that the results are within a margin / tolerance and slightly deviate from those results of the original study as long as the main claims in the original paper are not changed. This is especially true for non-computational studies (e. g., qualitative studies). It is not the responsibility of the reviewers to completely replicate / reproduce the study by themselves but of the authors to reasonably convey how this has been achieved. The goal of the Artifact Evaluation track is to promote work that allows the broader community to use the artifacts, not in-house specialists only. In case of Reusable artifacts emerging from, inter alia, more restrictive industrial research environments, the abstract needs to contain more than unreproducible claims of the artifact being used, i. e., sufficient details on the actual reproduction / replication to convince the well-intended reviewers.

5   Further Supplementary Material to this Document

While there are various (valuable) contributions related to open science and, thus, related to this guideline, we recommend the following supplementary material. Note that the guideline at hands is intended to be self-contained and the supplementary material is dedicated to the reader interested in the general notion of open science.

A broader introduction into the general notion of open science in Software Engineering, in particular open data and open source which we consider particularly important to the Artifact Evaluation Track, can be found in the (open access) book chapter "Open Science in Software Engineering", available here: . This chapter contains the ABC of open science and pragmatic, short insights into relevant basics such as proper licensing models.

The recommendations provided in the chapter are also reflected in the FSE Open Science policy, which we recommend to both reviewers and authors alike participating in the artifact evaluation track. See also the SIGSOFT Open Science Policies.

Finally, we recommend the general checklist elaborated by the Empirical Software Engineering research community as the ACM SIGSOFT Empirical Standards for researchers, peer reviewers, editors and publications venues.