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 artifact evaluation (AE) at scientific conferences. Since its initiation in the SE community at ESEC/FSE 2011, both the goals and the process of AE have evolved and today expectations towards AE are strongly linked with reproducible research results and reusable tools that other researchers can build their work on. However, to date little evidence has been provided that artifacts which have passed AE actually live up to these high expectations, i.e., to which degree AE processes contribute to AE's goals and whether the overhead they impose is justified.

We aim to fill this gap by providing an in-depth analysis of research artifacts from a decade of software engineering (SE) and programming languages (PL) conferences, based on which we reflect on the goals and mechanisms of AE in our community. In summary, our analyses (1) suggest that articles with artifacts do not generally have better visibility in the community, (2) provide evidence how evaluated and not evaluated artifacts differ with respect to different quality criteria, and (3) highlight opportunities for further improving AE processes.

CCS CONCEPTS

• **General and reference** → **Empirical studies**; • **Software and its engineering** → *Software post-development issues*; • **Information systems** → **Digital libraries and archives**.

ESEC/FSE '22, November 14–18, 2022, Singapore, Singapore

ACM ISBN 978-1-4503-9413-0/22/11.
https://doi.org/10.1145/3540250.3549172

KEYWORDS

Research artifacts, Artifact evaluation, Open science, Reproduction, Reuse

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 [2]. 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 communitities 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 [12, 13, 21]. In

[1] http://web.archive.org/web/20201031164603/http://2011.esec-fse.org/cfp-artifact-evaluation

particular, the tensions between high availability vs. high quality of artifacts and between the only partially overlapping goals of reproducibility and reusability are still being explored. Through the lens of reproducibility, artifact evaluation is a process centered on validating research results by reproducing those results using the artifacts supplied by the authors. Through the lens of reusability, artifact evaluation is a process centered on ensuring that artifacts will be publicly available and could be re-used and extended by future researchers.

What is clear, however, is that participation in artifact evaluation has grown enormously since its inception. It has been adopted at all major conferences and is increasingly adopted at journals in software-engineering and programming-language research. Adoption among authors has also increased over time. For instance, ≈ 90% of eligible papers at PLDI 2020 were accompanied by artifacts. However, this is not always the case for all venues.

The central question of our paper is: How can we, as a community, learn from our experiences in our first 10 years of artifact evaluation in order to improve the next 10 years? To gain corresponding insights, we inspect **(RQ1)** if articles accompanied by artifacts are more visible than those without, **(RQ2)** whether artifacts that passed evaluation are more often available, **(RQ3)** maintained after publication, **(RQ4)** more often reused, and **(RQ5)** more throughly documented. To inspect these aspects, we study conferences from the software engineering and programming language domains based on the selection made by Hermann, Winter, and Siemund [12] but limit our study to those where ACM guidelines apply to allow for a comparable baseline. We study the entire set of publications from these conferences in the past decade and identify artifacts which passed artifact evaluation but also those linked to a publication without a documented artifact evaluation badge.

From these insight we derive several suggestions how the artifact evaluation process may be improved in the inspected communities.

The contributions of our paper are:

(1) An in-depth analysis of how artifact evaluation practices impact paper and artifact outcomes, including both participation and quality
(2) Data-driven insights to improve artifact evaluation
(3) A dataset and associated tooling used to collect it to inspire further investigation or reproduction

2 BACKGROUND AND RELATED WORK

Background and Definitions. Claims in scientific literature must be supported by evidence or a reasoning why readers should regard the claims as valid [6]. Such evidence or reasoning in computer science research is often provided using a prototypical implementation, a collected or derived dataset, or an (automated) proof. Authors may choose to make these objects available (e.g., in the Archive of Formal Proofs[2]) for other researchers to inspect or reuse. The lack of availability of this supporting evidence has often been criticized to hinder reproducibility of research [17, 20].

A *supplementing artifact* is "a digital object that was either created by the authors to be used as part of the study or generated by the experiment itself. For example, artifacts can be software systems, scripts used to run experiments, input datasets, raw data collected in the experiment, or scripts used to analyze results" (ACM Task Force on Data, Software, and Reproducibility in Publication [8]). In this paper, we use the short term *artifact* to refer to such a supplementing digital object.

Artifact evaluation is the process of evaluating certain quality attributes of an artifact [12, 14, 15]. Typically, the evaluation work is done by an artifact-evaluation committee, which assesses whether artifacts are *reusable*, *functional*, *well-documented*, *consistent*, and *complete*.

An *artifact badge* is a pictogram to be displayed on a scientific article to declare quality attributes for a published research article. The first artifact badge in the PL community (Fig. 1) was introduced in 2013 for OOPSLA by Steve Blackburn and Matthias Hauswirth, and the properties to be evaluated were *easy-to-reuse*, *well-documented*, *consistent*, and *complete*.[3] It is still used for artifact evaluation in non-ACM conferences. Later, the ACM Task Force on Data, Software, and Reproducibility in Publication [4] introduced five colored badges to distinguish five different properties of artifacts.[5] The purpose of a badge is to reward artifact sharing and motivate authors to participate in artifact evaluation.

Figure 1: First badge from OOPSLA 2013

The five badges can be divided into three categories: (a) an artifact is *available*, independent from artifact evaluation, (b) artifacts satisfy the criteria of being *functional* or *reusable*, as assessed by artifact evaluation, and (c) results of the paper were *reproduced* with the artifact, or *replicated* without the artifact.

Figure 2: ACM badges

Artifact-evaluation committees are concerned with two or three of the above badges (*functional*, *reusable*, sometimes also *reproducible*[6]), while the *available* badge does not require evaluation (only that artifacts are long-term available, immutable, identifiable), and the *replicated* badge requires an independent study. There are different communities working on establishing standard processes and notions for badging of artifacts [18].

Related Studies. The expectations of the community regarding artifacts and their evaluation process were studied by Hermann, Winter, and Siegmund using a survey involving members from past artifact-evaluation committees [12]. The study raises several questions, some of which we strive to answer in this work. We particularly pick up on the quality aspect of artifacts and the effect of artifact evaluation on artifact quality. Heumüller et al. gave evidence that one of the most important expectations—the availability

[2] https://www.isa-afp.org/

[3] http://web.archive.org/web/20160217185935/http://evaluate.inf.usi.ch/artifacts/aea/badge

[4] http://web.archive.org/web/20211102201129/http://www.acm.org/publications/task-force-on-data-software-and-reproducibility

[5] http://web.archive.org/web/20220313070430/http://www.acm.org/publications/policies/artifact-review-and-badging-current

[6] It is debated in the community whether the *reproducible* badge requires an independent study or if it can be achieved through artifact-evaluation review.

of the artifacts described in scientific articles—is fulfilled only to an unsatisfactory degree [13]. In contrast to our study, they found a small positive correlation between linking to artifacts and citations to the article. However, they only inspected research track papers from the International Conference on Software Engineering (ICSE) in years without an established artifact evaluation process. Timperley et al. and Wacharamanotham et al. identified reasons for the insufficient availability of artifacts, and present a number of challenges that the authors encounter [21, 24]. A study on repeatability in computer-systems research reported that even if the artifacts are available, study results are often not reproducible [6]. The fields of computer systems [6, 10], computer graphics [5], communications [1, 26], and machine learning [11, 16] have also been the subject of studies on artifact quality and availability.

Data Collections. To ensure that artifacts are identifiable and findable, the relations between articles and artifacts must be reliably tracked and made available. Zenodo[7] provides a convenient interface to view, query, and change the relations of a digital object stored at Zenodo's digital library to other digital objects and provides means to declare the semantics of the link (such as 'is supplemented by this upload', 'is replaced by this upload', and 'cites this upload'). ACM's digital library has individual landing pages for artifacts and makes the links between articles and artifacts explicit. Article-artifact relationships that were found in a repeatability study [6] were made publicly available.[8] Baldassarre et al. collect reuse relationships between publication and artifacts beyond repeatability and reproduction [3]

3 RESEARCH QUESTIONS AND STUDY SUBJECTS

Our study addresses five research questions related to the merits of AE for authors, the merits of AE for artifact users, and how these merits for authors and users are linked with AE and publication processes and practices.

RQ 1: Are articles with artifacts that have passed AE more visible?
RQ 2: Are successfully evaluated artifacts more available?
RQ 3: Is artifact development/maintenance continued more often for successfully evaluated artifacts?
RQ 4: Are successfully evaluated artifacts more often reused?
RQ 5: Are successfully evaluated artifacts more thoroughly documented?

Before we discuss the relevance of these questions for the software engineering community, our methodology for answering these questions, and the results in more detail, we introduce the dataset through which the questions are investigated.

3.1 Subjects and Descriptive Statistics

The questions in our study are related to the effects of artifact evaluations. Therefore, we choose conferences from the SE and PL domains that have implemented corresponding processes. Hermann, Winter, and Siegmund [12] provide a comprehensive list of such conferences that we use for our subject selection. As the format and degree of information that conferences provide regarding the conducted AE differs significantly, we restrict our study to conferences with proceedings in the ACM Digital Library (DL). The main reason for this decision is that the ACM's guidelines for artifact review and badging [8] provide a common, albeit very general, AE framework and that all AE processes adopting this framework should be comparable on that basis. Moreover, the ACM DL provides uniform formats for (1) proceedings, (2) publication metadata, and (3) research artifacts linked with publications, which facilitates the creation of a consistent dataset.

Figure 3 shows the conferences with AE for which we have collected article data from ACM's DL by conference and year. We refer to the combination of conference and year as *venue*. We had to exclude FSE 2012 and MODELS 2019 from our dataset. For FSE 2012, only a "best artifact award" was awarded by the program committee. The number of candidates for this award or the selection process remain confidential. Hence, we were not able to identify which articles had *evaluated* artifacts that were considered for the award and which had not. Therefore, we cannot make any meaningful comparison between artifacts that did and did not undergo an evaluation. For MODELS 2019, only the workshop papers from the companion proceedings are available in the ACM DL. However, there is no information regarding any AE process or evaluated artifacts for these workshops available. We added ASE 2018 to our dataset, although it did not have a formal artifact evaluation process, because "Available" badges were issued for some of the articles and we can, therefore, assess effects that we attribute to badges (rather than AE processes) as targeted by RQ1. On the top of each bar in Figure 3, a number indicates the total number of articles in our dataset. This number may be smaller than the actual number of articles in the proceedings, as we exclude keynotes, workshop abstracts, etc. More precisely, we include every article from the proceedings that has an author, is tagged as "Research Article" or "Article" in the ACM DL, and has at least a total length of 4 PDF pages. This collection does include short papers, as for several venues tool papers are short papers that may have undergone AE and, hence, are relevant for our study. In total, our analysis in the following sections is based on 3650 articles from 64 venues. The bars in Figure 3 also indicate the relative fractions of different article categories relevant to our study.

4 RQ1: ARE ARTICLES WITH ARTIFACTS THAT HAVE PASSED AE MORE VISIBLE?

Preparing artifacts for AE entails significant amounts of work for authors. However, evaluation metrics for hiring, promotion, and tenure often are centered on the visibility of articles — derivatives of publication and citation counts — not on the visibility of artifacts. While we do not endorse the use of these metrics for evaluating career advancement, it is nonetheless the case that many institutions around the world rely on them, and some researchers are forced to optimize towards them. One hypothesis is that AE positively impacts the visibility of publications [13]. If this hypothesis holds, it may provide authors with a strong incentive to participate in AE. If it does not, an investigation of alternative reward mechanisms may be worthwhile.

[7] https://zenodo.org
[8] http://www.findresearch.org/

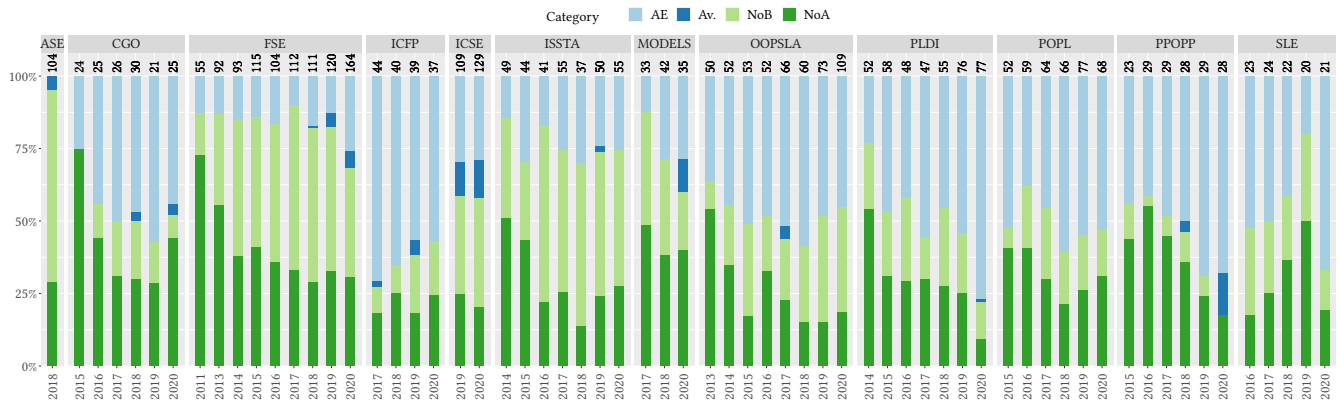

Figure 3: Percentage of articles per category across all venues in our study. The numbers on top of the bars display the total number of articles for the venue. For an explanation of categories, refer to Section 4.1

4.1 Method

We measure visibility in terms of citation counts of articles, which we obtain from Crossref [7].

To link these visibility measures with AE, we group articles into four article categories: **(AE)** With artifact & AE badge ("Functional",[9] "Reusable",[10] or the "old venue-specific badges" (Figure 1)), **(Av.)** with artifact and only the "Artifact Available" badge, **(NoB)** with artifact but without any badge, or **(NoA)** without artifact. Note that we treat articles that only have an "Artifacts Available" badge separately from articles that also have other badges as "Artifacts Available" does not imply an actual evaluation of the artifact [8], as discussed for ASE 2018 above.

We identify categories (AE) and (Av.) by their badges in the ACM DL. To identify the old monochrome badges, which are not shown in DL article entries, we extract the upper left and right corners from article PDFs and analyze the distribution of pixel colors in those areas to detect the presence of a badge. To rule out false positive matches due to irregular formatting, we manually confirmed each badge detection. To rule out false negatives, we compared the number of detected badges against the number of accepted artifacts reported on conference websites and (if available) to information provided by Conference Publishing [19] and the http://www.findresearch.org portal. We additionally consulted the artifact evaluation chairs' reports in proceeding front matters and contacted the AEC chairs of the conferences for confirmation.

We identify categories (NoB) and (NoA) by conducting a tool-assisted manual review and classification of 25 728 URLs from 3150 article PDFs (the remaining PDFs in our dataset did not contain any URLs). In this process, we automatically extract URLs from the PDF text and manually tag each extracted URL as "accessible artifact URL", "inaccessible artifact URL", or "no artifact URL". We make the tool available together with our dataset in the artifact accompanying this paper [25].

To determine whether AE affects visibility, we determine whether there is stochastic dominance of either category over any other category by conducting a two-sample Kolmogorov-Smirnov (KS) test on each pairwise combination of the four categories. We perform these pairwise tests separately for each conference and year to avoid effects from "age" on visibility metrics [13]. To account for confounding factors, we further categorize papers by (a) page lengths, and (b) whether they are published as public or closed access. In addition to these confounding factors, we also attempted to analyze confounding with article topic as per the 2012 ACM Computing Classification System (CCS). However, we found the spread of CCS topics to be too large to support a meaningful analysis, e.g., the FSE 2020 proceedings feature 75 different CCS categories that only apply to one single paper, 22 that apply to 2, 13 to 3, 4 to 4 and so on. Consequently, a stratification of the dataset according to CCS categories would lead to numerous strata with single or few articles and, thus, impede a meaningful comparison. We, hence, decided to exclude the impact analysis for CCS categories from our study of confounding factors, but kept the data in our artifact [25].

Our data does not meet the prerequisites for parametric approaches to confounding control (e.g., citation counts do not follow a normal distribution). Other approaches like multiple linear regression or logistic regression assume a linear relationship between the independent variables and the dependent variable (respectively, its logit). We, thus, analyze the impact of these variables on the association between article categories and citation counts by stratifying our data accordingly and analyzing differences in citation counts across all strata using KS tests. As we test for each potential confounding factor (page lengths and open/closed access) and their combinations, we conduct a total of 16 KS tests ($(2 + 3!) \cdot 2$, as we test for both directions of possible stochastic dominance) per venue and adjust our p values accordingly for multiple testing using the Benjamini-Hochberg (BH) procedure [4]. Based on the outcome of these tests, we conduct 12 KS tests against stratified or unstratified data from each of the four categories (we compare each of the four categories against the others) and perform correction on the obtained p values as before.

[9]For the colored ACM badges, we consider both versions 1.0 and 1.1.
[10]For PPoPP 2020 and CGO 2020, "Results Replicated" badges were issued in the AE and we, thus, consider them as well for these conferences.

4.2 Results

Table 1 shows the conferences and article categories for which citation count distributions are statistically significantly ($\alpha = 0.05$) affected by differences in the identified confounding variables "page length" (regular vs. short papers, where we set the cut-off at 10 pages) and "open/closed access" (OA/CA). The columns list the results for the strata for which we identified stochastic dominance relations, indicated by >. We find significant effects of confounding variables in 14/64 venues. For all of them, regular papers have significantly higher citation counts than short papers and the relation for CA Reg. > CA Short likely is a direct effect of that. For the other potential confounding factors, there is no clear pattern.

Table 1: Statistically significant ($\alpha = 0.05$, after BH correction) effects of article page counts (distinguishing Short from Reg. papers), open/closed access (OA/CA), and their combinations on citation counts.

Venue	Reg. > Short	CA Reg. > CA Short	OA Reg. > CA Short	CA Reg. > OA Short	OA > CA
ASE 2018	✓	✓	–	–	–
FSE 2011	✓	✓	–	–	–
FSE 2013	✓	✓	–	–	–
FSE 2014	✓	✓	–	–	–
FSE 2015	✓	✓	✓	–	–
FSE 2016	✓	✓	–	✓	–
FSE 2017	✓	✓	✓	–	✓
FSE 2018	✓	✓	✓	–	–
FSE 2019	✓	✓	✓	–	–
FSE 2020	✓	✓	✓	–	–
ISSTA 2015	✓	✓	–	–	–
ISSTA 2017	✓	✓	✓	–	–
ISSTA 2018	✓	✓	–	–	–
ISSTA 2019	✓	✓	✓	–	✓

We subsequently stratify the citation data according to the levels of the confounding variables (i.e., open vs. closed access and page counts less vs. greater than or equal to 10 pages) for the conferences, for which we found a significant effect of these variables (indicated by the tick marks in Table 1) and conduct our analysis on the respective strata. The results (p values and KS statistic D as effect size measure) are shown for statistically significant cases ($\alpha = 0.05$) in Table 2. After BH correction, we only find statistically significant citation count differences between (NoB) and (NoA) short papers published at FSE 2014 and FSE 2019.

Contrary to other analyses [13], our results indicate that articles with artifacts do not generally get more citations. We are only able to confirm statistically significant effects for 2 out of 64 venues in our study. Moreover, the significant effects we observe are limited to articles without badges (NoB) and to the short papers category. Therefore, we conclude that creating and publishing research artifacts does not generally have beneficial effects on citation counts.

Finding 1: Artifacts do not significantly improve citation counts of research articles.

Table 2: p values and KS statistic D (in braces) for statistically significant ($\alpha = 0.05$, after BH correction) effects of article categories on citation counts. ">" indicates which category has a significantly greater citation count. Strata are indicated in braces in the venue column.

Venue	NoB > NoA
FSE 2014 (CAShort)	0.013 (0.562)
FSE 2014 (Short)	0.013 (0.562)
FSE 2019 (Short)	0.047 (0.492)

5 RQ2: ARE SUCCESSFULLY EVALUATED ARTIFACTS MORE AVAILABLE?

The reproducibility of research results and the reusability of research artifacts are perceived as the main objectives of artifact evaluations by AEC members [12]. If an artifact is not available, it can neither be reused, nor can the paper results be reproduced. Therefore, availability is a vital quality criterion for artifacts.

5.1 Method

To study, whether artifacts that passed AE are more often available than artifacts that did not, we classify artifacts as *(a) passed AE* (article group (AE) in RQ1) or *(b) unknown* (article groups (Av.), (NoB) and (NoA) in RQ1). We refer to these groups as AE and NonAE in the following.

To test whether an artifact from either group is available requires at least three steps.

(1) There must be an artifact reference, e.g., as a URL in a published article.
(2) The artifact reference must be resolvable to one or more digital objects (e.g., downloadable files or web services).
(3) The referenced digital object must be an artifact of the paper as per the definition in Section 2.

Testing for the second criterion can be automated (with bounded precision), whereas testing for the first and third requires manual investigation.

(1) Artifact Reference Availability: To identify whether a research artifact reference is available for articles in our study, we search different information sources for these references:

The ACM Digital Library (DL) [9] provides authors of published articles with the opportunity to also publish any accompanying research artifacts. Artifacts in the DL have their own dedicated records with links from and to the research articles that they accompany.

Conference Publishing is a consulting agency for publishers of scientific articles. Conference Publishing is entrusted with the publication processes for a large number of SE and PL conferences and openly publishes metadata for these conferences on its website [19]. Artifact links from Conference Publishing are extracted as the "info links" that author can supply when submitting their camera ready article versions.

findresearch.org is a platform that presents semi-automatically collected metadata of computer science research articles. Authors

of research articles are queried for confirmation of presumably automatically extracted data[11]. The portal does not contain metadata for conferences after 2018, but serves as a reference for older venues in our study.

CMU dataset: In a recent study of research artifacts [21], the authors have manually analyzed artifact references in research articles and published this data [22]. As the venues covered by that dataset overlap with the venues in our study, we make use of the dataset for intersecting venues.

Article PDF files: For the venues in our study that are not covered by the dataset at [22], we conduct a similar analysis as the authors in [21]. As the manual analysis of PDF URLs does not scale well for the total of 3650 articles in our study, we have developed a tool ("URLBrowser") to support this process. The tool automatically extracts URLs from PDF files and opens these links in a web browser to facilitate URL classification (whether the URL points to an artifact of the paper and, if so, whether that link works). We make URLBrowser publicly available as part of our artifact [25].

(2) Digital Object Availability: To approximate the availability of digital objects referenced by the URLs identified in the first step of our availability analysis, we send HTTP HEAD requests using cURL [23] and analyze the returned HTTP status codes. This measurement only yields an approximation, because (a) websites may be available, but not contain the artifact (false positives) and (b) websites may not respond to HEAD requests (false negatives). On a manually investigated sample of 200 links that were flagged as available (the sample discussed in Section 8) and 416 links that were flagged as unavailable, we found 2.5 % of false positives and 5.6 % of false negatives. In addition to this automated process, we utilize results from the analysis of article URLs using URLBrowser, as detailed above.

(3) Correspondence of Available Digital Objects to Research Artifacts: Whether an available digital object qualifies as a research artifact is non-trivial and one of the central questions targeted by artifact evaluations. An in-depth analysis of all 3685 digital objects, for which the cURL-based analysis indicated availability, is not manageable within the scope of this article. We, thus, rely on the AEC's assessment for artifacts that underwent AE (article group AE). For artifacts from other article groups, we rely on the assumption that the manual investigation of the digital object's reference with URLBrowser is sufficiently indicative of whether the digital object is indeed an artifact of the analyzed research article.

5.2 Results

The availability results from the outlined procedure are shown in Table 3. The table is partitioned into AE and NonAE articles and further divides these partitions based on whether the article carries an "Available"badge. This information is relevant, because artifacts may have undergone AE but not been made publicly accessible. Similarly, if we were not able to find an artifact reference for an article without a badge, that may either mean that there is no artifact for this article or that we were not able to find its reference. We can only distinguish between those cases for articles carrying an "Available"badge. The last three columns list for each of the four resulting partitions the number of articles, the number of articles with a reference, and the number of articles with at least one accessible reference.

Table 3: Accessibility of artifacts with AE badge and Av. Badge. Note: AE indicates artifact was evaluated. NonAE could indicate the authors did not submit artifact for evaluation, or they did and the AE committee did not award a badge. Percentages are calculated based on the neighboring column to the left.

AE Evaluated	Available Badge Status	Total Papers	Has Artifact Reference	Is Accessible
AE	Av. Badge	683	676 (99.0%)	675 (99.9%)
	No Av. Badge	602	473 (78.6%)	431 (91.1%)
NonAE	Av. Badge	71	67 (94.4%)	65 (97.0%)
	No Av. Badge	2294	1148 (50.0%)	1032 (89.9%)

Reference Availability: A comparison of the first two numerical columns reveals that we were not able to identify artifact references for 11 articles with "Available"badges, 7 of which also carry AE badges. A closer inspection of these cases reveals that 2 cases of articles with AE badges and all 4 of the articles without are publications at ICSE 2020. As AE for this venue followed an open review process, the artifacts are indeed available in the GitHub repository[12] on which the review process was based. Unfortunately, the repository is only linked from the submission information page for the venue and not in the publication itself or any publication metadata available in common databases for scientific literature. Three of the remaining 4 cases are due to insufficiencies in our URL detection: One of them is a reference to a privately hosted git server (which is no longer accessible, but the link is provided in the article), one is due to font encoding issues in the article's PDF file (which also affects other text-based functions, such as searching text in the article), and one is missed by the PDF to text conversion underlying our URLBrowser tool for unknown reasons. One of the remaining 2 articles makes an unspecific reference to the ACM DL, but we were not able to find further information there. However, we were able to find GitHub repositories for the 2 artifacts via a web search. In summary, we were not able to identify references for 8 out of 754 analyzed articles with an "Available"badge from the published text or publication metadata and only found them via a web search or looking into the details of the artifact submission and management process for the venue.

Digital Object Availability: A comparison between the second and third numerical column in Table 3 shows how many references returned failure-indicating HTTP status codes upon an attempt to access the referenced digital objects. The numbers reveal differences, both between articles with and without "Available"badges and between articles with and without AE badges. For only 3 articles with an "Available"badge (1 with and 2 without AE badges), we could not find any working reference among the articles' references, which accounts for 0.4 % of the articles with "Available"badges. In

[11] We were not able to identify the precise source of artifact links on findresearch.org

[12] https://github.com/researchart/rose6icse/tree/master/submissions/available

Table 4: Artifact references in our study by host platform. The first column lists the type of host platform, the following the number of total and broken references and their ratio.

Reference Type	Found AE	Broken AE	% Broken all Data	Broken NonAE	Found NonAE
IP Address	2	2	100%	1	1
Other	2	2	100%	1	1
File Storage	10	5	36%	4	15
Web Application	14	1	25%	5	10
URL Redirection	18	4	24%	16	64
Institutional Website	362	56	23%	114	371
Company Website	14	1	21%	7	24
Project Website	84	10	14%	25	164
Personal Website	27	3	12%	5	39
Public Archive	66	1	5%	6	78
Public VCS	749	26	4%	37	778
Publisher Aux. Material	754	11	1%	1	78
DOI/Handle	182	1	1%	2	74
Youtube	3	0	0%	0	75

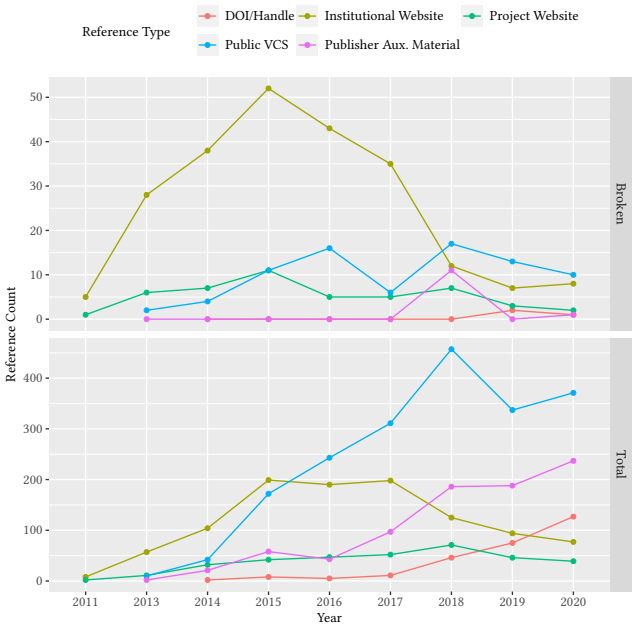

Figure 4: Total and broken artifact links according to HTTP response for different platforms.

contrast, for articles without an "Available"badge that number is 158 (9.7 %). While we could not find any working reference for 43 (3.7 %) articles with AE badges, the number for articles without AE badges is 118 (9.7 %).

For the four broken references from articles with "Available"badges, we manually investigated the cases and found two of the references (one in the AE group, one in the without AE group) to be falsely identified as not working by our HTTP status code based detection. We do not consider such false detections to affect our overall conclusion from the presented data due to the large difference between article groups with/without "Available"/AE badges.

These results indicate that the overall number of papers with references to research artifacts is similar across the two partitions AE (1149 articles) and Non-AE (1215). However, "Available"badges, which are associated with low fractions of broken references, are much more prevalent in the AE partition.

As the "Available"badge has only been introduced to AE with ACM's standardization of artifact badges in 2017, there is a possible confounding of the observed effect with reference age. As the central criterion for awarding the "Available" badge is that the artifact is hosted on a platform with a long retention policy, we analyze the effects of host platforms on artifact availability and which hosting platforms have been most prevalent over time. We identify host platforms by extracting the domain of a given artifact reference and manually classifying it as, for instance, institutional websites, personal websites, project websites, public version control systems (VCS), etc.

All 4071 artifact references in our study can be classified according to the 14 link categories listed in Table 4. The left side of the table lists the number of references for each category in AE articles and the number of broken references identified by our cURL-based check. The right side of the table lists the same information for NonAE articles. The table rows are ordered by the overall fraction of broken to total references ("% Broken all Data"). Besides YouTube, which possibly contains false positives as the site shows a custom error page and does not return HTTP 404 on missing content, we see that in particular DOI/handle links and publisher auxiliary material (e.g., artifacts hosted in ACM's DL) have a low broken reference ratio and significantly more AE than NonAE references fall into these categories. Moreover, both categories fulfill the long-term retention requirements for the "Available" badge.

Figure 4 displays how the numbers from Table 4 distribute over time. To maintain visibility, we only display host platforms with at least 50 links in at least one year. while the upper part of the figure shows the number of broken references by platform and year, the lower part shows the total number of references as a baseline. From the figure we see that a large number of broken references point to institutional websites. We also see in the lower part that from 2017, the number of references pointing to institutional websites or project websites decreases. The references to publisher auxiliary material and DOIs/Handles increases, while the number of broken references to these platforms remains low. While the steep decrease of broken institutional website links between 2017 and 2018 must be partially attributed to recency, as the drop of total references in that category is somewhat smoother (albeit on a different scale), we do expect the observed change in publication culture due to the requirements set forth by the "Available"badge to have a lasting impact due to the long-term retention they mandate.

Finding 2: Due to the hosting platform requirements they entail, "Available" badges are positively linked with artifact availability.

6 RQ3: IS ARTIFACT DEVELOPMENT/MAINTENANCE CONTINUED MORE OFTEN FOR SUCCESSFULLY EVALUATED ARTIFACTS?

If an artifact continues to be maintained and developed, that indicates that it is reused and, therefore, must have been reusable and is/was of high quality, at least for the period of maintenance/development.

6.1 Method

To measure development and maintenance activity, we rely on information from public version control systems. As most articles in the (AE) and (NoB) classes provide GiHub links, we focus our analysis on GitHub and use its REST API to obtain the following information: (1) the time of the last commit, (2) the number of commits after artifact publication, (3) the number of contributors, (4) the number of forks, and (5) the number of stars/watchers.

The first two measures are the central measures for answering the RQ, as we use them to calculate (a) the development time period after artifact publication ("Dev. Time"), (b) the time between the last commit and the date of our data collection ("Idle Time"), and (c) the number of commits during Dev. Time ("Commit Density"). We use the other three metrics as indicators of interest and visibility of the artifacts.

6.1.1 Results. Table 5 shows the results of the KS tests we conducted to assess the difference in GitHub-based metrics for the development/maintenance activity after artifact publication. The tests are based on data from a total of 1920 repositories (900 belonging to AE articles and 1020 to NonAE articles). p values are not adjusted, as only two tests (one for each direction of possible stochastic dominance) are conducted for each of the disjoint metrics.

AE repositories have significantly higher commit density and dev. time in addition to a significantly shorter idle time. This indicates that these repositories are indeed used for the active development of AE artifacts, even beyond their submission to artifact evaluation, and not for archiving them. For NonAE, the lower development activity indicates that authors mainly use the repositories for artifact archival. This impression is strengthened by the generally higher interest and visibility metrics (contributor, star, and watch counts), which may either indicate usage as "bookmarks" or hope for further evolution of the projects (which does not seem to occur in the general case). Fork counts are also significantly different between AE and NonAE repositories, but without clear stochastic dominance of either group over the other. The KS statistic D, which indicates the maximal percentage difference between the two groups' cumulative distribution functions, is moderate with a maximal difference of 15.4 % (for star counts) across the metrics.

Finding 3: Repository-based activity, interest, and visibility metrics are higher for evaluated artifacts.

Table 5: p values and KS statistic D for statistically significant ($\alpha = 0.05$) differences in GitHub statistics based on article categories (AE: with AE badges, NonAE: without AE badges). p values are not adjusted, as only two tests per metric are conducted. Stochastic dominance is indicated by ">".

Metric	p	D
Idle Time (NonAE > AE)	<0.01	0.086
Dev. Time (AE > NonAE)	<0.01	0.085
Commit Density (AE > NonAE)	<0.01	0.112
Contributor Counts (AE > NonAE)	<0.01	0.088
Star Counts (AE > NonAE)	<0.01	0.154
Watcher Counts (AE > NonAE)	<0.01	0.142
Fork Counts (AE > NonAE)	<0.01	0.093
Fork Counts (NonAE > AE)	<0.01	0.076

7 RQ4: ARE SUCCESSFULLY EVALUATED ARTIFACTS MORE OFTEN REUSED?

Availability of research artifacts is a necessary, but not a sufficient prerequisite for their utility to reuse in scientific research and result reproduction. To serve the research community, artifacts must also be reusable for reproducing research results or for repurposing in different contexts. We, therefore, analyze how often artifacts are being reused.

7.1 Method

We analyze references to research artifacts to approximate reuse. If an artifact is referenced in a research article, that indicates that the artifact has been useful for other work. To analyze referral to artifacts, we search for the presence of artifact links (obtained from various sources as discussed for the *Availability* quality criterion above) within article PDFs. We restrict our search to the articles in our dataset and use the URLs extracted in the article classification process for RQ1 (see Section 4.1). Our URL matching accounts for small differences that do not affect the identity of the digital object being referenced (presence/absence of trailing slashes or a "www." prefix). We include references from years *before* the artifact's discussion in a publication, as the artifact may have been available and useful before an article discussing the related research has been accepted for publication. As referral by others than the original authors of the artifact indicates better reusability (others must be assumed to be less familiar with the artifact's usage and structure), we also take the overlap of author groups between the referring article and the referenced article discussing the artifact into account.

7.2 Results

Table 6 shows the absolute numbers and relative fractions of articles with referenced artifacts and referencing articles in our study. The first column indicates whether we count articles with or without intersecting author lists. To make a comparison between references to AE and NonAE artifacts, we partition our dataset and the second column of the table labels the rows accordingly. The numbers yield different conclusions depending on whether author lists do or do not intersect. For articles with intersecting author lists, more NonAE

Table 6: Articles in our study that are referenced by and that are referencing other articles by artifact URL without overlap in referring/referred author groups.

Author lists intersect	Category (Referenced)	Referenced Articles	Referencing Articles
yes	AE	69 (0.05%)	80 (0.06%)
	NonAE	83 (0.07%)	96 (0.08%)
no	AE	48 (0.04%)	61 (0.05%)
	NonAE	40 (0.03%)	138 (0.11%)

artifacts than AE artifacts are referenced, whereas the opposite is true for articles with non-intersecting author lists.

We also see more references (last column) to NonAE artifacts, irrespective of whether author lists intersect or not. The difference between references to NonAE versus AE papers is larger for non-intersecting author lists (138 vs. 61) than for intersecting author lists (96 vs. 80). This means that while slightly fewer NonAE than AE artifacts are referenced in our dataset (column 3), they are referenced in more articles (column 4).

In summary, we cannot draw clear conclusions from the data. The higher number of referenced AE artifacts by non-intersecting author groups may be seen as a weak indication that evaluated artifacts are easier to reuse by authors that were previously unfamiliar with the artifact. But at the same time, the smaller number of referenced NonAE artifacts is referenced by a larger number of articles. Our results are limited to articles in our dataset and we plan to extend our analysis to a larger corpus of articles in future work.

Finding 4: More AE artifact links are being referenced, but more references exist to the fewer NonAE artifacts.

8 RQ5: ARE SUCCESSFULLY EVALUATED ARTIFACTS MORE THOROUGHLY DOCUMENTED?

According to published results, documentation is perceived as an important quality criterion for artifacts by many users and past AEC members [12, 21]. If an artifact contains no or little documentation, it is difficult to reuse and the paper results are likely difficult to reproduce, which clearly limits its quality.

8.1 Method

As the analysis of documentation requires the download of linked artifacts, which requires a manual investigation of the linked web sites, we restrict our analysis to a randomly drawn sample of 100 artifacts for each category (AE and NonAE). As each article may contain multiple artifact links from different sources, we prioritize links from the ACM DL over links found in PDF files over links from other sources, i.e., Conference Publishing or findresearch.org. The rationale for this prioritization is that links in PDF files are provided by authors and easily identified by readers. The links published by Conference Publishing and findresearch.org are also author submitted links, but are usually not directly visible to readers, unless they explicitly search for information on these platforms. To address the imbalance of archive file types (e.g., zip or tar) vs. repository links in the ACM DL compared to other sources, we generally give preference to archive file types over other links from the same source. In the case of links to Zenodo, we use Zenodo's REST API to resolve the artifact link to download links of files linked with the Zenodo record. For git repository links, we attempt to checkout the versions that got accepted/rejected during AE, where we determine the date as the AE notification date if that information is available. If that information was not available, we either used the camera-ready due date (if indicated as relevant for artifacts as well on the venue's website) or the date of the venue's program announcement.

To approximate the adequacy of artifact documentation, we search for document file types in the artifact and quantify the amount of documentation by word counts. As we are not aware of any existing standards for research artifact documentation or widely accepted practices, we then proceeded to search for 12 (case insensitive) file name patterns across the identified document files according to our experience with research artifacts "^read.*me", "^setup", "^install", "^doc/", "^examples?/", "^assets?/", "^artifact", "detailed.*result.*pdf", "report.*\.pdf", "supplement.*\.pdf", "^copyright", "^license". The first four items target typical file names with initial instructions for software projects. The next three keywords are inspired by our observation that research artifacts we have evaluated or worked with contain them and that artifact-related information of larger projects is kept in dedicated artifact directories. The next three keywords target detailed technical documentation extending the published article. Finally, the last two keywords indicate the presence of licensing information, which is a mandatory prerequisite for (re-)use of the artifact.

8.2 Results

Table 7 shows the result of our documentation analysis of 100 sampled AE and 100 sampled NonAE artifacts. In our randomly drawn sample of 100 artifacts each, 13 artifacts in the AE group and 12 artifacts in the NonAE group did not include any file matching any of our search terms. For the artifacts with missing documentation, there is no pattern in terms of conference or year. Out of the remaining artifacts, 84 from the AE sample and 86 from the NonAE sample contained a README file with a much higher average word count for files in the AE sample. At the same time, separate documentation and examples directories are more common and their content is more comprehensive (in terms of average word count) for the NonAE sample than the AE sample. A deeper analysis of the collected data reveals that the large amount of documentation and examples in the analyzed NonAE sample is only contributed by a comparatively small fraction of 15 artifacts, out of which only 5 exceed the single observed AE "^doc/" word count and only 2 the mean AE "^examples?/" word count. We see this as an indicator that the artifacts in the NonAE sample are highly diverse with only few artifacts providing extensive documentation.

Overall, the results show that AE artifacts tend to have more comprehensive overview documentation in README files, whereas we observe some NonAE artifacts to have a shorter overview documentation and a more comprehensive documentation in separate

directories, albeit in a limited number of cases. While the former is more suitable for a focused reproduction of research results, the latter is more suitable for repurposing and reuse, which may hint at underlying differences in the perceived purpose of research artifacts [12]. The documentation of positively evaluated artifacts focuses on reproduction, whereas the documentation of (some) other artifacts focuses on reuse and repurposing.

Based on the observation that a top level README file is missing for 15% of the sampled artifacts and our difficulties to identify suitable search terms for artifact documentation, we furthermore recommend the development of community standards for artifact packaging and evaluation. At FSE, for instance, certain documentation is now required to be included with the artifact submission and we recommend to develop similar unified community-wide standards for artifact submissions, evaluation, and archival.

From our results, we also see that the majority of the sampled artifacts do not seem to contain proper licensing information. As some of the analyzed artifacts were obtained from the ACM DL and Zenodo, which provide means to specify artifact licenses on the artifact web pages, we additionally investigated the presence of licensing information for artifacts on those platforms, whenever we did not find a license file.

Out of the artifacts for which we did not find a license, 26 artifacts from the AE sample and one artifact from the NonAE sample were hosted on the ACM DL or Zenodo. For 22 of those from the AE sample, we were able to obtain the license, whereas we were not able to obtain any license for 4 artifacts from the AE sample and the NonAE artifact. We investigated these cases manually to confirm the absence of licenses. All 4 artifacts from the AE sample are hosted on the ACM DL, which, contrary to Zenodo, does not strictly mandate a license specification. The artifact from the NonAE sample is hosted on Zenodo and the authors chose the "Other" option for the license, which is commonly used if different parts of the artifact are published under different licenses. This indeed is the case for the artifact, but for some parts the license files are missing, which leaves the terms of use for those parts unclear.

Out of the 26 artifacts with licenses specified in the publication metadata, 4 were also available on other platforms, which did not include a license file. Therefore, we generally consider it advisable to include license files with the artifacts, rather than just in the metadata.

Finding 5: Documentation practices strongly differ across AE and NonAE artifacts and within the NonAE group. Many AE and NonAE artifacts are lacking licenses and copyright information.

9 THREATS TO VALIDITY

The chosen methodology to answer our research questions results in a number of threats to the validity of our conclusions.

Construct validity: Participation in artifact evaluation, as a variable of interest, is not directly measurable because AE processes and review are not typically public. Our categorization focuses on artifact badges as an indicator, because the corresponding papers are known to have passed AE. Based on a limited set of open review based AE processes (FSE 2016 & 2018, ICSE 2020, MODELS 2018, and SLE 2016), we assume the number of artifacts that may have benefited from the AE process despite being rejected is negligible.

Table 7: Number of articles with file names matching the given search terms. Word counts are averages across the given numbers of articles and rounded to integer values.

Search Term	Matched Artifacts		Word Count	
	AE	NonAE	AE	NonAE
No match	13	12	–	–
^read.*me	84	86	1,389	645
^install	6	1	324	593
^doc/	1	8	2,431	13,901
^examples?/	4	9	1,470	426,353
^assets?/	1	1	10,412	657
^artifact	6	1	2,973	1,203
report.*\.pdf	1	1	1,822	42,789
supplement.*\.pdf	1	1	2,222	2,086
^copyright	0	1	0	268
^license	50	46	850	1,220

Internal validity: Except for RQ2, the findings of our study are based on associations between article categories and other variables and do not hypothesize causal relations. For RQ2, we detail why we consider a causal relation between hosting platform and artifact availability reasonable.

We control for confounding factors in our analysis to the degree possible by the data that is available to us. Especially in terms of how AE is conducted, how AEC chairs implement/guide an AE process may have a strong effect and we cannot trivially assess that, because it is rarely documented.

To control for confounding, we stratify our dataset according to hypothesized confounding factors, test for differences across them, and maintain the stratification if the observed effects are significant (after correction for multiple testing). The resulting stratification leads to smaller sample sizes within the strata, which reduces the discriminative power of the subsequent tests for differences in our response variable (citations). As a consequence, the reported results are conservative.

To control for selection bias, we include a wide range of SE and PL conferences that adopted artifact evaluation over several years using different processes, from which we randomly sample artifacts for our documentation analysis. The selection of subjects in our study is restricted to conferences organized or supported by the ACM, for which publication and artifact data is available through the ACM Digital Library. We have taken great care to analyze potential effects of this choice, e.g., by cross-comparing the obtained metadata with other sources, e.g., from Conference Publishing. During our consistency checks, we identified and reported a number of data inconsistencies to ACM, which got acknowledged and fixed.

Our approach of identifying artifact URLs from PDFs is imperfect and represents a threat to internal validity. We mitigate that threat by (a) providing the set of identified artifact URLs as part of our dataset, allowing them to be scrutinized and (b) including

our tooling for identifying artifact URLs as part of our replication package.

External validity: Our analysis and, thus, our conclusions are limited to the SE and PL venues, for which proceedings are available in the ACM DL. However, this sample accounts for 64 out of 89 datapoints (i.e., almost 71.9 %) according to the most comprehensive study of AE adoption in the SE and PL communities to date [12].

10 DISCUSSION

In the first decade after its initiation, artifact evaluations have significantly gained popularity in the SE and PL communities. In our article, we look at the artifacts evaluated during this period and make a comparative assessment with research artifacts that have not been submitted or did not successfully pass artifact evaluation. In this section, we discuss our findings and make recommendations to further improve artifact evaluations for the coming decade.

AE Reward Mechanisms: The main reward mechanism for artifact submitters are badges, which are prominently displayed in the title area of articles and in digital libraries like the ACM DL. However, in our study we find that this advertisement of research results obtained with evaluated artifacts does not significantly affect the visibility of research articles in terms of citations. As much of the traditional academic performance evaluation is centered around citation-derived metrics, the creation and maintenance of artifacts is currently not well integrated in this system, also because there is no standardized way to reference them and they are, hence, likely to escape the common citation tracking mechanisms. As the creation of these artifacts entails significant overheads, we encourage the SE and PL communities to propose and discuss alternative reward mechanisms for authors who create and publish high quality research artifacts, which benefit the research community as a whole. Besides more rigorous attribution policies for artifacts, which could be included in peer review guidelines, alternative reward mechanisms could also be based on non-citation metrics, e.g., the number of positively evaluated artifacts or the R+ index [3]. With a decade of artifact publications, the addition of test-of-time awards for artifacts may also reward creators of particularly useful artifacts and constitute a valuable addition to conference programs.

"Available" Assessment: In Section 5 we discuss the impact the "Available"-badge-imposed requirements have on the availability of research artifacts. However, from Figure 3, we see that there is an up and down in AE participation and that (NoB) articles still dominate for SE conferences (see FSE, ISSTA, ICSE), even after 2017, when the "Available" badges were introduced. For PL conferences the situation is a bit better, but there is generally very little reason to not get "Available" badges for any (NoB) article. We suspect that the reason for this partially is that the process for obtaining "Available" badges is often linked with the artifact evaluation process. Authors, who do not want to get an actual evaluation of their artifacts may not be aware of the "Available" badge option. At the same time, we have seen some "Available" articles in our dataset, for which we could not easily find links. This could be prevented by introducing an additional check for the camera-ready version of articles whether they contain an artifact reference if they are assigned the "Available" badge.[13] In summary, we recommend to link the "Available" badge assignment with calls for papers and the paper review process, rather than the artifact evaluation. We also recommend to focus further research on the factors that prevent authors from packaging, submitting, and publishing their research artifacts, as we expect the related insights to significantly benefit our communities' processes and the availability of research artifacts.

Community Standards: Our analysis of the documentation for a sample of artifacts has revealed deficiencies regarding the presence of common documentation and license files. This means that such information is either indeed missing or that it is hidden in places not covered by our analysis. To make sure that this information is present for every artifact and that it can be easily found, we recommend the communities to develop common standards for the documentation of artiacts. FSE, for instance, is currently mandating certain information to be present in certain files in the artifact submission and we endorse to adopt and extend this standardization effort. Specifically, standards could also cover apsects of artifact packaging, submission, publication, and referencing, which would facilitate artifact reviews as well as automated checks to scale with the hopefully further increasing numbers of artifact publications in the coming years.

DECLARATIONS

Data-Availability Statement. All data and scripts are available on Zenodo [25].

Funding Statement. This work is supported by the National Science Foundation under Grants 2100037 and 2100015. Open access was funded by the LMUexcellent Fund.

ACKNOWLEDGMENTS

We thank ACM, Conference Publishing, and the many AEC chairs from past SE and PL venues for their help with filling gaps in our dataset. We appreciate the anonymous reviewers' feedback and constructive suggestions for improving our manuscript.

REFERENCES

[1] Vaibhav Bajpai, Anna Brunstrom, Anja Feldmann, Wolfgang Kellerer, Aiko Pras, Henning Schulzrinne, Georgios Smaragdakis, Matthias Wählisch, and Klaus Wehrle. 2019. The Dagstuhl Beginners Guide to Reproducibility for Experimental Networking Research. *SIGCOMM Comput. Commun. Rev.* 49, 1 (feb 2019), 24–30. https://doi.org/10.1145/3314212.3314217

[2] M. Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533 (May 2016), 452–454. https://doi.org/10.1038/533452a

[3] Maria Teresa Baldassarre, Neil A. Ernst, Ben Hermann, Tim Menzies, and Rahul Yedida. 2021. Crowdsourcing the State of the Art(ifacts). *CoRR* abs/2108.06821 (2021). arXiv:2108.06821 https://arxiv.org/abs/2108.06821

[4] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

[5] Nicolas Bonneel, David Coeurjolly, Julie Digne, and Nicolas Mellado. 2020. Code Replicability in Computer Graphics. *ACM Trans. Graph.* 39, 4, Article 93 (jul 2020), 8 pages. https://doi.org/10.1145/3386569.3392413

[6] Christian S. Collberg and Todd A. Proebsting. 2016. Repeatability in computer systems research. *Commun. ACM* 59, 3 (2016), 62–69. https://doi.org/10.1145/2812803

[7] Crossref. 2022. Crossref Metadata Search. https://search.crossref.org/. Accessed: 2022-03-14.

[13] Conference Publishing has decided to follow this recommendation and a corresponding check will already be implemented for FSE 2022.

[8] Association for Computing Machinery. 2020. Artifact Review and Badging — Current. https://www.acm.org/publications/policies/artifact-review-badging-current. Accessed: 2021-08-27.
[9] Association for Computing Machinery. 2022. ACM Digital Library. https://dl.acm.org/. Accessed: 2022-03-17.
[10] Eitan Frachtenberg. 2022. Research artifacts and citations in computer systems papers. *PeerJ Computer Science* 8 (Feb. 2022), e887. https://doi.org/10.7717/peerj-cs.887
[11] Odd Erik Gundersen, Saeid Shamsaliei, and Richard Juul Isdahl. 2022. Do machine learning platforms provide out-of-the-box reproducibility? *Future Generation Computer Systems* 126 (2022), 34–47. https://doi.org/10.1016/j.future.2021.06.014
[12] 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)*. ACM, New York, NY, USA, 469–480. https://doi.org/10.1145/3368089.3409767
[13] Robert Heumüller, Sebastian Nielebock, Jacob Krüger, and Frank Ortmeier. 2020. Publish or perish, but do not forget your software artifacts. *Empir. Softw. Eng.* 25, 6 (2020), 4585–4616. https://doi.org/10.1007/s10664-020-09851-6
[14] Shriram Krishnamurthi. 2013. Artifact evaluation for software conferences. *ACM SIGSOFT Softw. Eng. Notes* 38, 3 (2013), 7–10. https://doi.org/10.1145/2464526.2464530
[15] Shriram Krishnamurthi and Jan Vitek. 2015. The real software crisis: Repeatability as a core value. *Commun. ACM* 58, 3 (2015), 34–36. https://doi.org/10.1145/2658987
[16] Chao Liu, Cuiyun Gao, Xin Xia, David Lo, John Grundy, and Xiaohu Yang. 2021. On the Reproducibility and Replicability of Deep Learning in Software Engineering. *ACM Trans. Softw. Eng. Methodol.* 31, 1, Article 15 (oct 2021), 46 pages. https://doi.org/10.1145/3477535
[17] National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303
[18] NISO. 2021. *Reproducibility Badging and Definitions: A Recommended Practice of the National Information Standards Organization*. Technical Report NISO RP-31-2021. https://doi.org/10.3789/niso-rp-31-2021
[19] Conference Publishing. 2022. Conference Tables of Contents and Programs. https://www.conference-publishing.com/. Accessed: 2021-08-29.
[20] Martin Shepperd, Nemitari Ajienka, and Steve Counsell. 2018. The role and value of replication in empirical software engineering results. *Inf. Softw. Technol.* 99 (2018), 120–132. https://doi.org/10.1016/j.infsof.2018.01.006
[21] Christopher Steven Timperley, Lauren Herckis, Claire Le Goues, and Michael Hilton. 2021. Understanding and Improving Artifact Sharing in Software Engineering Research. *Empir. Softw. Eng.* 26, 4 (2021), 67. https://doi.org/10.1007/s10664-021-09973-5
[22] Christopher S. Timperley, Lauren Herckis, Claire Le Goues, and Michael Hilton. 2021. *Replication Package for Understanding and Improving Artifact Sharing in Software Engineering Research*. https://doi.org/10.5281/zenodo.4737346
[23] Contributors to the cURL project. 2022. cURL: command line tool and library for transferring data with URLs. https://curl.se/. Accessed: 2022-03-17.
[24] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. *Transparency of CHI Research Artifacts: Results of a Self-Reported Survey*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376448
[25] Stefan Winter, Christopher S. Timperley, Ben Hermann, Jürgen Cito, Jonathan Bell, Michael Hilton, and Dirk Beyer. 2022. Reproduction Package (Docker Container) for the FSE 2022 Article 'A Retrospective Study of One Decade of Artifact Evaluations'. Zenodo. https://doi.org/10.5281/zenodo.7082407
[26] Noa Zilberman and Andrew W. Moore. 2020. Thoughts about Artifact Badging. *SIGCOMM Comput. Commun. Rev.* 50, 2 (may 2020), 60–63. https://doi.org/10.1145/3402413.3402422