The Journal of Systems and Software 218 (2024) 112187

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

Research artifacts for human-oriented experiments in software engineering: An ACM badges-driven structure proposal ☆

Cathy Guevara-Vega [a,b,*], Beatriz Bernárdez [b,c], Margarita Cruz [b], Amador Durán [b,c], Antonio Ruiz-Cortés [b,c], Martin Solari [d]

[a] eCIER Research Group, Universidad Técnica del Norte, Ibarra, Ecuador
[b] I3US Research Institute, Universidad de Sevilla, Seville, Spain
[c] SCORE Lab, Universidad de Sevilla, Seville, Spain
[d] Universidad ORT Uruguay, Uruguay

ARTICLE INFO

Keywords:
Research artifact
Laboratory package
Replication package
Empirical software engineering
Human-oriented experiments
Open science
ACM (Association for Computing Machinery)
Badging policy

ABSTRACT

Context: The Open Science (OS) movement promotes the value of making public the research artifacts (datasets, analysis scripts, guidelines, etc.) used during empirical studies. OS is widely known in areas such as Medicine or Biology, where the process of sharing research artifacts is subject to strict protocols. Unfortunately, in Software Engineering (SE), this process is carried out in a non-systematic way, resulting in incomplete or inaccurate material shared by researchers, which hinders the reproducibility and replicability of empirical studies. Nevertheless, in recent years, it seems that the Empirical Software Engineering (ESE) community is embracing some of the proposed OS initiatives, such as the one proposed by the Association for Computing Machinery (ACM), which provides a badge system to evaluate the quality of a research artifact. This badge system has been adopted by several SE conferences as a method of assessing research artifacts.

Aims: Focusing on human-oriented experiments (HOEs) in SE, whose research artifacts are more complex than those for computational experiments, this work applies Design Science Research (DSR) with a twofold purpose: (i) review the current status of HOEs research artifacts publication through evaluation of this practice in the most relevant ESE journals , and (ii) propose a structured outline for HOEs research artifacts driven by the aforementioned ACM badging policy.

Method: Regarding the first purpose, we carried out a survey to analyze the current status of the publication of research artifacts considering relevant peer review journals and the quality of 106 research artifacts published in these journals with respect to the ACM badging policy. For the second purpose, an iterative process was carried out to review the content of 106 research artifacts research and their concordance with ACM badges, obtaining a structured scheme for HOEs research artifacts that has been validated through a detailed review of 12 research artifacts obtained from some of those of ACM badges in relevant SE conferences. In addition, we validated the proposal in the research artifacts of 2 of our own experiments.

Results: Our survey reveals issues such as the 39,70% of journal studies making completely accessible their research artifacts; most of the analyzed research artifacts are incomplete; the most common repositories used in the ESE community to share the research artifacts are GitHub, institutional repositories, and Zenodo. On the other hand, the validated and structured research artifact outline consists of a list of ordered sections containing a set of artifacts, which can be mandatory or not to achieve a particular ACM badge. For its internal validation, several improvement iterations on the first release of the outline have been carried out based on the conference guidelines, the ACM badging policy, and other relevant proposals.

Conclusions: Although the ESE community is making great efforts in standardization, review, and digital publishing related to OS, the availability and completeness of research artifacts can be improved. Our proposal for the elaboration of structured research artifact outline meets the requirements of HOEs in SE. Nevertheless, further research is needed not only to improve and externally validate it but also to disseminate its use among the research community.

☆ Editor: Prof Neil Ernst.

* Corresponding author at: I3US Research Institute, Universidad de Sevilla, Seville, Spain.
E-mail addresses: cguevara@utn.edu.ec (C. Guevara-Vega), beat@us.es (B. Bernárdez), cruz@us.es (M. Cruz), amador@us.es (A. Durán), aruiz@us.es (A. Ruiz-Cortés), martin.solari@ort.edu.uy (M. Solari).

https://doi.org/10.1016/j.jss.2024.112187
Received 31 July 2023; Received in revised form 11 August 2024; Accepted 16 August 2024
Available online 27 August 2024

1. Introduction

Open Science (OS) is a movement that promotes sharing the data from empirical studies, the material used to analyze the data, the source code, and the manuscripts reporting on the studies, i.e., *open data*, *open material*, *open source*, and *open access*, respectively, in publicly available repositories (Méndez et al., 2020). As UNESCO (United Nations Educational, Scientific and Cultural Organization), highlighted in a comprehensive report on OS (UNESCO, 2021a), the potential provided by digital communication technologies and global interconnectedness is a key factor in accelerating scientific progress and fostering knowledge societies.

In this regard, OS initiatives from the empirical research community are intended to make scientific studies more transparent and encourage their replicability to verify the results and produce a solid theory (Popper, 1959). The fact that the data – and the rest of the artifacts – are available for the community fosters reanalysis and reliability of the results (Gómez et al., 2010), in addition to new authors being able to replicate the study by having access to the material used for further querying, reusing, and repurposing (ACM, 2020). In summary, the goal of OS is to increase the *transparency, reproducibility* and *replicability* of empirical studies (Méndez et al., 2019), while they reduce the tacit knowledge (Juristo and Gómez, 2010; Shull et al., 2002).

In this context, a *research artifact* is defined by ACM (2020) as a digital object created by the researchers to be used as a part of a study or generated by an experiment itself. Therefore, a research artifact or simply an *artifact* comprises the collection of elements or files that have been generated throughout the planning, completion, analysis, and reporting of a given empirical study. For some time now, several proposals encouraging the development of research artifacts have been presented by relevant authors from the Empirical Software Engineering (ESE) community, as the works by Basili et al. (1996, 1999), as well as *The PROMISE Software Engineering Repository*, born in 2005 (as a part of an ICSE workshop) in order to initiate the generation of a publicly available repository of Software Engineering (SE) datasets (Sayyad Shirabad and Menzies, 2005). One of the most recent approaches is Solari's et al. which provides: (i) the description of a series of modules that set a good research artifact content (Solari et al., 2018); and (ii) the definition of five research artifact maturity levels related to the research artifact quality (Solari and Vegas, 2006).

Apart from the development of research artifacts and its free availability, another movement to promote OS is the concept of *badge*, a system of incentives in which a research artifact receives one or more recognitions according to a peer quality review. Among these, the most established badges in the ESE community are those developed by the Association for Computing Machinery (ACM, 2020) and the proposal of Open Software Foundation (OSF – Center for Open Science) (Blohowiak et al., 2022).

As a result of the aforementioned proposals, SE community has pioneered the concept of *artifact evaluation* (AE), and some conferences have recently included an AE track. This track debuted during ESEC/FSE 2011 (Winter et al., 2022) and has since been adopted by major SE conferences such as ASE, RE, and ICSE. As part of the AE track, authors submit their research artifacts, and experts review them to grant some of the aforementioned badges. While the primary goal of AE tracks is to assess, endorse, disseminate, and organize research artifacts, typically through the allocation of ACM badges, a recent study conducted among AE committee members (Hermann et al., 2020) emphasized the absence of uniform quality standards and clear criteria according to which artifacts must be awarded. This was attributed to the broad scope of the ACM guidelines, which allows for diverse interpretations.

Another important initiative in favor of OS is the emergence of several repositories where authors can host their research artifacts persistently and free of charge, in contrast to uploading materials to personal websites, which are more error-prone and may become unavailable.

Despite all these initiatives, the current literature reveals that developing a good research artifact is a challenge for the authors, who have reported issues such as a lack of standards and guidelines, hosting and maintenance, and the feeling that developing a good artifact is not worth it (Timperley et al., 2021).

After carefully studying the OS-related proposals mentioned above, we have adopted Design Science Research (DSR) as our research methodology with minor changes (Peffers et al., 2007). Fig. 1 shows the complete picture of the whole study across the iterative DSR phases pursued to address the following *overarching questions*.

Phase (I). *What are the main proposals for OS initiatives? What is the landscape in ESE regarding developing research artifacts?* In this phase, we carefully analyzed the related bibliography to identify problems and research niches, as well as motivate our work (Sections 1 and 2).

Phase (II). *What are the publication trends of research artifacts, guides, and repositories used by researchers? Which are the main issues that researchers found when they adopt OS initiatives?* In this phase, we carried out a survey (by carefully reviewing 106 research artifacts) to identify and motivate the issues when the ESE community has embraced the initiatives mentioned above and applied them to their works on human-oriented experiments (HOEs) (Wohlin et al., 2012) (Sections 3 and 4). After the survey, we realized the usefulness of providing a structure for HOEs according to ACM badges.

Phase (III). *Which should the structure and content of a complete quality research artifact based on Solari et al. proposals (Solari et al., 2018) and ACM guidelines (ACM, 2020)?* In this phase, we provided a structure for HOEs research artifacts based on such proposals. Following an iterative process of design and development, we were able to produce the first version of the outline that was analyzed and reviewed with the selected journal research artifacts (Section 5).

Phase (IV). *How good is the proposal in terms of coverage and adequacy for research artifacts of HOEs?* In this phase, we compared the content of 12 awarded artifacts – taken from AE tracks of four reference SE conferences – with our proposal, based on the ACM badge level achieved at the conferences as an award of the corresponding AE track. We also carried out an internal validation of the proposal against research artifacts from 2 experiments conducted by some of the current authors (Sections 5.1 and 6 for threats to the validity of the whole study).

As a result, we have obtained a structured outline for HOEs research artifacts that comprise the structure of a computational experiment research artifact, which is usually not so complex. Note that computational experiments research artifacts usually include *packages installed in the language*, *libraries installed in the host system*, *file names*, and *directory hierarchy* (Wonsil, 2021; Trisovic et al., 2022).

The 106 research artifacts belong to studies published in the journals EMSE, IST, JSS, and TSE,[1] due to these journals (i) have been considered the most suitable journals for reporting HOEs in SE (Revoredo et al., 2021); (ii) have incorporated OS considerations in their *guide for authors*; and (iii) are included in the top ten of the ranking of leading journals in SE & Programming (Research.com, 2023). Table 1 summarizes the main properties of these journals related to the OS initiative. EMSE was the first journal to incorporate the initiative at 2019, and all forums promote open material. Only EMSE has its own badging system (EMSE journal, 2023).

[1] Empirical Software Engineering Journal; Information and Software Technology Journal; Journal of Systems and Software Journal; and IEEE Transactions on Software Engineering Journal.

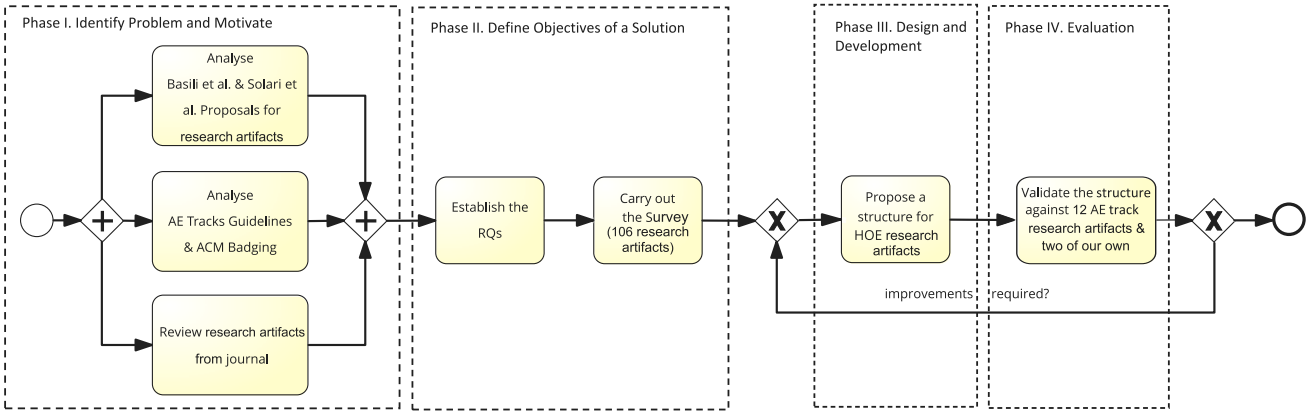

Fig. 1. BPMN diagram for the iterative process followed to develop a HOEs research artifact structured outline.

Table 1
Open science in journals under study.

Journal	Start year of OS	Open data	Open material	Open software	Open access	Badging
EMSE	2019	✓	✓	✓	✓	OS Badge
TSE	2020	✓	✓	✓	✓	–
IST	2021	–	✓	–	✓	–
JSS	2021	–	✓	–	✓	–

2. Background

2.1. Open science

As commented in the previous section, OS fosters the initiative to make any research artifact available to the scientific community, including data, materials, source code, and free access to articles and reports. OS also encourages sharing of preprints, infrastructure details, licensing models, author's rights, and review models such as double-blind peer review (Graziotin, 2021). OS is a means of working together with the scientific community, fostering research networks, and co-creating knowledge.

OS is addressed by expert groups such as CERN (2023) and UNESCO (2021b), academic communities (FORCE11, 2021) and book publishing (Miedema, 2022; Nielsen, 2012; Weinberger, 2014). In particular, the report published by UNESCO (2021a) provides (i) a definition for OS and (ii) a framework for developing OS initiatives at different levels.

(i) *OS sets a new paradigm that integrates into the scientific enterprise practices for reproducibility, transparency, sharing, and collaboration resulting from the increased opening of scientific contents, tools, and processes.*

(ii) *Open scientific knowledge refers to open access to scientific publications, research data, metadata, open educational resources, software, and source code and hardware that are available in the public domain or under copyright and licensed under an open license.*

After the aforementioned recommendations, frequently applied in other mature sciences such as Medicine, Biology, and Psychology (Dasgupta and Paul, 1994; Munafò et al., 2017; Woelfle et al., 2011), OS is gathering momentum in the SE scientific community, in which researchers have exhibited remarkable acceptance and enthusiasm for disseminating their research artifacts. On the one hand, some of the main SE conferences such as ICSE, ESEC/FSE, ASE, and RE have recently included a new AE track whose *call for artifacts* promotes the submission of research artifacts. Then, an expert committee reviews research artifacts and grants some of the ACM badges.

In this context, peer-reviewed journals have recently adopted OS initiatives focusing on various aspects of the scientific process to make science more transparent, reproducible, and accessible. Open-access publication models have evolved according to the requirements of the funding agencies. Authors have self-archiving options at several stages and can enable public access to their manuscripts. Journals also encourage the publication of research data and associated artifacts. With the aim of enabling reproducible research, Elsevier has proposed a framework for data management processes and systems throughout the data lifecycle (Elsevier, 2023). Many major SE journals have OS initiatives. Journals like EMSE, IST, JSS and TSE encourage authors to submit a research artifact to an open repository such as ArXiv, Zenodo, or GitHub.

EMSE (2023) prompts and highlights the OS initiative to make research data public and increase the *transparency, reproducibility*, and *replicability* of the published studies. This supports the immediate credibility of the authors' work and provides an everyday basis for joint community efforts based on shared data. In addition, it builds a general body of knowledge in the community that leads to SE theories for applicability (Méndez et al., 2019).

IST (2023) and JSS (2023) support OS initiative by providing their authors with a free preprint publication service on SSRN (Social Science Research Network), an Open Access Preprint Online Community. SSRN serves as a platform that enables early registration and dissemination of preliminary research findings. This facilitates early citation and collaboration among researchers (SSRN Elsevier, 2022).

In addition, the TSE (2023) journal supports the chapter of the *2020 Ad Hoc Committee on OS and Reproducibility* (Open Science IEEE Computer Society, 2020), to analyze models, practices, and experiences in supporting OS and Reproducibility within the IEEE Computer Society and at peer societies and publishers in the context of the recommendation by NASEM (2022).

EMSE, IST, and JSS journals have implemented a separate review process for research artifacts and linked datasets (EMSE journal, 2023; IST Journal, 2023; JSS Journal, 2023). IEEE Computer Society has created an *ad–hoc* Committee on OS and Reproducibility and has implemented an infrastructure for dataset sharing called IEEE DataPort (IEEE, 2023). Similarly, JSS has not only a specific OS Board (JSS Journal, 2023) but also an agreement with Mendeley, by which there is a specific repository for the datasets of articles published in JSS (Mendeley Data, 2023).

2.2. ACM badging policy

The ACM (2020) presents three groups of badges that artifacts from empirical studies can obtain when the research artifact is submitted to a conference AE track or to a special issue of OS in journals, each characterized as follows.

Evaluated Evaluated is defined based on a set of quality characteristics that the research artifact should accomplish, such as

completeness or structure. The *Evaluated* category is divided into two different badges. The *Functional* badge indicates that the corresponding research artifact is documented, consistent, complete, and exercisable. The *Reusable* badge adds additional quality characteristics, such as being well-structured and well-documented, to those in the *Functional* badge.

Available is awarded if the research artifact is shared in a repository with a DOI, and relevant artifacts are provided there.

Validated is awarded to an original experiment when it was replicated by researchers independent of the ones that carried out the original study. *Validated* is also divided into two different badges, *Reproduced* and *Replicated*. Since the original experiment has been externally replicated, it could be awarded with *Reproduced* badge if the external replication was carried out using, in part, the experimental setting of the original experiment. On the other hand, it could be awarded. The *Replicated* badge when the external replication used its own experimental setting (such as a different measurement system or experimental material).

Due to the considerable abstraction inherent in ACM badges (ACM, 2020), when a conference incorporates an AE track, it requires that specialists elucidate the badges, taking into account the particular themes pertinent to the conference's research domain, and formulate guidelines to construct a meritorious research artifact eligible for recognition. These guidelines include suggestions such as (i) an inventory of the artifacts and a description is required; (ii) The artifacts should be relevant to the study and add value beyond the text in the article; (iii) 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; or (iv) If a research artifact is for installation, software artifacts, scripts, and source code should be documented along with instructions on obtaining them.

2.3. Research artifact

To the best of our knowledge, the beginning of the development of research artifacts in SE dates back to the mid-1990s, with the publication of experimental material by Kamsties and Lott (1995a) in an internal report of the ISERN.[2] (Kamsties and Lott, 1995b) After that, under the name of *laboratory package*, Basili et al. (1996, 1999) proposed (i) the first approach for the summary composition of a good research artifact, (ii) the idea of the artifact availability on the Internet and (iii) the development of an insightful research artifact for a family of experiments on requirement verification techniques (Basili et al., 2006). Although some authors adopted this term, such as Shull et al. (2002), Pucci et al. (2014) and Solari et al. (2018), several related terms for research artifact are identified in the scientific community. In Conradi et al. (2001) and Mian et al. (2005) the authors define it as *experimental package*, in Heumüller et al. (2020), Robillard et al. (2018), Roper et al. (1997) and Solari and Vegas (2006) as *replication package*, in Abrahão et al. (2019) and Runeson et al. (2014) as *experimental material* and in Santos et al. (2021) as *supplementary material*. In this work, we will use the term *research artifact* due to the SE community having settled on this term.

A research artifact is a set of elements that must be structured, validated, and available for knowledge transfer concerning a *family of experiments* (Santos et al., 2018). As commented above, research artifacts should support all the activities in the experimental process, promoting *replicability*, *transparency*, and *reproducibility* of empirical studies (Solari, 2011).

[2] International Software Engineering Research Network

The importance of conducting replications to consolidate the findings of empirical studies has been recognized in the scientific community (Guevara-Vega et al., 2021; Shepperd et al., 2018). Replication is a complex task by nature, as shown by Cruz et al. (2019); therefore, it is desirable to have research artifacts that are complete and easy to understand De Magalhães and Da Silva (2013). The information provided in a research artifact must be useful for knowledge transfer among researchers (Solari et al., 2018) thus, this communication with the original experimenters favors replication success (Vegas et al., 2006; Shull et al., 2008). Brooks et al. (1994) also point out that without a research artifact, it is unlikely to verify an experiment by replication.

Since then, some authors have made experimental material – datasets, statistical analysis, or questionnaires – available in various forms. For instance, research artifacts, appendices, supplementary material for the corresponding article, often on personal websites (Almqvist, 2006). Nevertheless, according to Solari et al. (2018), research artifacts are not necessarily limited to the exchange of experimental material; a good research artifact may include descriptions, a replication process guide, and procedures to avoid tacit knowledge and to encourage further replications (Brooks et al., 1994; Basili et al., 1996). ACM SIGSOFT has also highlighted the importance of sharing research artifacts in its report *ACM Empirical Standard* (Ralph et al., 2020), which presents a set of checklists to assess the completeness of an empirical research article according to its content, emphasizing that making public the de-identified dataset is *essential* (when there are no incompatibilities for ethical reasons) and includes supplementary material, such as research protocol, task materials, or analyses scripts, is *desirable*. In the following subsections, we summarize several research artifact taxonomies, as they play a crucial role in facilitating a comprehensive discussion in the current study. Specifically, they contribute to answering $RQ_{2.2}$ and provide essential insights for our proposal of structure for HOE research artifacts.

2.3.1. Research artifact taxonomy regarding its maturity level

Solari and Vegas (2006) evaluate the maturity of research artifacts by proposing the following categories described below, from the lowest to the highest maturity level.

- *Article*. The scientific article is not accompanied by a research artifact and it is the only source that presents information about the experiment.
- *Operational*. The research artifact includes the artifacts that provide the software production components, instructions for the experimental subjects, forms, measurement instruments, conceptual models, UML diagrams, etc.
- *Descriptive*. The research artifact includes descriptions of the elements contained in the package for the replicator to understand and use them properly. The design used should be described, and the experiment context should be explained in detail.
- *Procedural*. The research artifact includes guidelines for the experimental process that can be followed thoroughly (step by step) by the replicators.
- *Decision-making*. The decisions made by the original experimenters to perform the experiment are justified. These rationalizations allow replicators to follow and understand the reasons that encourage them to conduct the experiment while making their own decisions. Hence, this leads to a reduction of tacit knowledge.

2.3.2. Research artifact taxonomy regarding the material shared with community

Solari et al. (2018) proposes seven modules that classify and delve into the content of a research artifact.

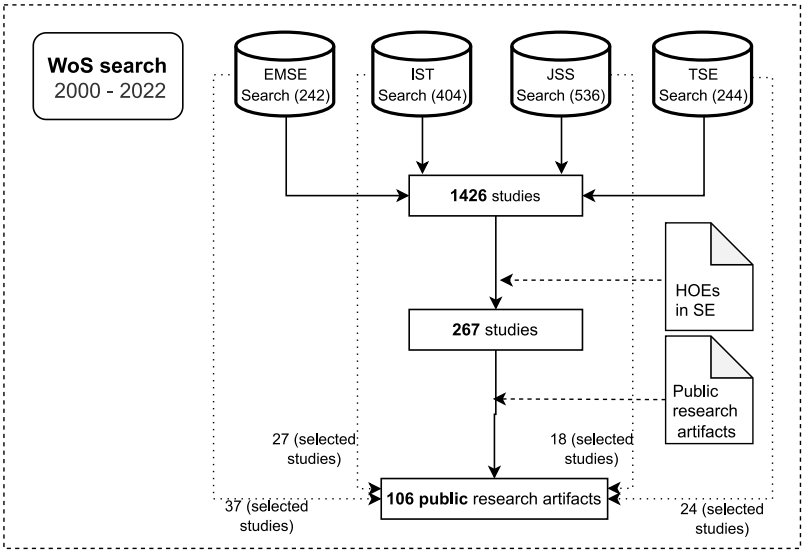

Fig. 2. Selection process for research artifacts in journals.

- *Introduction*. Contains essential information about the research artifact. It is designed to provide the replicator with information about the research artifact structure and use. In addition, it indicates the objective and context of the experiment, as well as the authors' data for carrying out a replication.
- *Theory*. Establishes the theoretical framework for conducting the experiment and interpreting the results, describing the techniques used in the experiment, bibliographic references, conceptual models, UML diagrams, and formulas.
- *Training*. Contains all materials used to train the experimental participants (if applicable), e.g., training guidelines, and feedback sessions.
- *Experiment*. Defines the experiment process and includes operating material (such as datasets or analysis scripts), instructions for replication, and experimental tools.
- *Evolution*. Records the activities that the experiment has undergone, namely the list of replications, aggregations, and versions of the experiment. This module establishes a relationship between the versions of the experiment and its replications to represent how the research artifacts evolve as a whole. As an example, when several replications of the investigation are performed, the experiment evolves, and the experimental objective changes (to generalize the results). This module should reflect the historical evolution of the experiment across the whole family.
- *Replication*. This module records information about each replication, including a description of the adapted design, the resulting data, and notes on the replication experience.
- *Aggregation*. Gathers analysis of data from several experiments that make up a family of experiments, with meta-analysis or other comparison techniques.

Although Solari's framework for artifact development is exhaustive and of superior quality, there remains an imperative need within the scholarly community for more detailed and precise guidelines pertaining to the development of research artifacts (Méndez et al., 2019; Hermann et al., 2020; Timperley et al., 2021). Moreover, the development of an initiative that not only meets Solari's criteria but also aligns with the ACM badges could enrich the domain of research artifact development.

3. Survey in OS philosophy in key journals

This section explains the process followed to select the research artifacts to conduct the survey in order to answer the following questions RQ_1 and RQ_2. We followed the guidelines proposed by Petersen et al. (2008) with minor changes.

RQ_1 *What is the current status of HOE research artifacts publication considering relevant peer review journals?*

RQ_2 *What findings on completeness are revealed by conducting a quality analysis following the previous ACM guidelines (ACM, 2020) and Solari's (Solari et al., 2018) proposals?*

The question RQ_1 was split into four more detailed level questions to point out specific goals.

$RQ_{1.1}$ *Is it a current practice in the ESE community to make HOE research artifacts public?*

$RQ_{1.2}$ *How is the number of published HOEs research artifacts evolving in the ESE community?*

$RQ_{1.3}$ *How do authors use the guidelines provided by journals to report their HOE research artifacts?*

$RQ_{1.4}$ *What digital repositories are being used to make HOE research artifacts available in the ESE community?*

Similarly, RQ_2 has been divided into the following sub-questions.

$RQ_{2.1}$ *What is the level of alignment of the research artifacts with respect to the ACM badging policy?*

$RQ_{2.2}$ *What is the level of completeness of the research artifacts concerning Solari's module proposal?*

3.1. Planning

The search strategy starts by querying the Web of Science (WoS) database repository with a custom search string. After several search attempts, we decided to use the term "experiment" to ensure high retrieval. We accepted the risk of including articles reporting any type of experiment with this query. To overcome low search precision, we performed a manual review of each scientific article. The inclusion criteria taken into account in the search were (i) studies that include the term "experiment" in the title or in the abstract; (ii) studies published since 2000; and (iii) studies published in EMSE, IST, JSS or TSE.

Journal studies research artifacts that did not present HOEs in SE were excluded.

3.2. Conduction of the search

The search was conducted by querying the WoS repository with the following query string:

((TI = (Experiment)) OR AB = (Experiment)) AND
(SO = (IEEE Transactions on Software Engineering) or
SO = (Journal of Systems and Software) or
SO = (Information and Software Technology) or
SO = (EMPIRICAL SOFTWARE ENGINEERING))
AND DOP > 1999

The query was executed on January 2023 and returned 1426 studies. After applying the exclusion criteria, 267 studies remained (see Fig. 2).

3.3. Study selection and classification

The exclusion criteria were applied mainly by one researcher to ensure that all the studies and research artifacts were reviewed uniformly. Only in case of doubt, a second researcher was consulted.

After applying this manual process, 267 journal studies, out of which 106 defined public research artifacts were selected and comprise our primary studies. Fig. 2 shows the research artifact selection process.

Finally, to analyze the set of 106 public research artifacts embraced in this survey, which is *available* in the supplemental material at *Zenodo* (Guevara-Vega et al., 2024), we defined categories such as quality level, package content, and type of access, among others, to classify the data of each research artifact to answer the RQs.

4. Results

In this section, we answer the research questions set out in Section 1, all of which are referred to the Empirical Software Engineering community.

4.1. Current practice to make research artifacts public ($RQ_{1.1}$)

To address this question, it should be noted that in journals, authors have the freedom to decide whether to publish their research artifacts or not, although editors often encourage them to do so.

Of the 267 journal studies analyzed, 147 defined a research artifact whereas 120 – almost half of the studies (44.94%) – did not. Also, 106 are publicly available, i.e. 39,70% (106/267) as can be seen in Fig. 2. Fig. 3 shows a stacked column chart of the 267 analyzed journal studies. For each journal, the percentage of studies without research artifacts and their availability when they contain research artifacts are shown. The four journals present a similar percentage of research artifacts with public URLs, although it is slightly higher in TSE and EMSE.

On the other hand, by focusing on 147 journal studies that defined a research artifact, we could identify three different scenarios, (i) 8 studies with research artifact that require access authorization (5,44%); (ii) 33 studies with research artifacts with broken URLs (22.44%); and (iii) 106 studies that provide research artifacts with publicly accessible URLs for retrieval amount (72,12%), which aligns with the best practice of following the OS philosophy.

In addition to the study of the availability of research artifacts, from the 267 journal studies, we have analyzed how the content of the research artifacts is described in the own scientific article. The authors usually split the information between several sections of the paper and one or more appendixes as supplemental material. As a result, in the paper, the authors explain not only the experimental design and process (relevant topics considered as part of a complete research artifact following Basili et al. (1999), but also the step-by-step guide on how the experimental tasks were carried out and the material used. On the other hand, in the appendix, survey formats, additional analyses, results of experimental subjects' experience, call graphs, training artifacts, measuring instruments, and conceptual models are occasionally

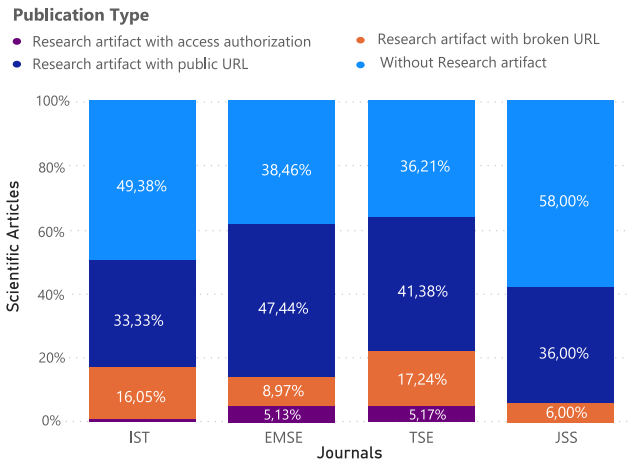

Fig. 3. Percentage of articles in the selected journals reporting a HOE and availability of their research artifacts.

provided. Either in the sections of the article or in the appendices, some authors include URLs to access the different artifacts, such as surveys, informed consent, readme files, licenses, scripts, and datasets. Finally, we have observed that the EMSE journal is the one that publishes the most articles that deepen the research artifact, which is usually described in the manuscript itself, with an extensive number of pages.

4.2. Trends in artifact publication($RQ_{1.2}$)

With this RQ, we want to know how the publication of journal studies including research artifacts has evolved between 2000 and 2022. The trend can be seen in Fig. 4.

The timeline is every 5 years, and the percentages are calculated according to the number of studies published each year. There is an average of 5 studies per year that publish a research artifact, so no significant variation is observed because the number of studies that publish research artifact is relatively low. However, in 2020, there was a considerable increase in 35 studies (instead of 5 studies as in previous years) publishing research artifacts in the four journals (EMSE, IST, JSS, TSE). One factor that may have influenced this high number is the OS initiative that emerged in previous years from the different scientific events, e.g., conferences have included track or sessions exclusively to evaluate research artifacts and boost the quality of research artifacts through the ACM badging policy, whose first version was published before 2018 (Ferro and Kelly, 2018).

4.3. Guidelines to report research artifacts ($RQ_{1.3}$)

With this question, we want to study the guidelines provided by journals so that experimenters can develop their own research artifacts.

After analyzing the guidelines of the forums studied in this article, Table 2 summarizes the conclusions of the calls for OS. It should be noted that: (i) the guidelines are very general; (ii) they are focused on *open–material*; (iii) journal publishers have web pages wherein they refer to the OS initiative and its relation to reproducibility; (iv) one of the journals – EMSE – works with badges and grants awards called the *OS badge*. While studying the white papers published by EMSE related to OS, we identified the intention to adopt the ACM badging system in the near future.

After carefully analyzing the 106 research artifacts, we observed they do not fully follow the recommendations proposed in the OS prescriptions of the corresponding journals.

In summary, assessing this question, we can check that the relevant authors have claimed that for replications of HOEs in the area of SE, the diversity of research and the lack of guidelines make the content of the

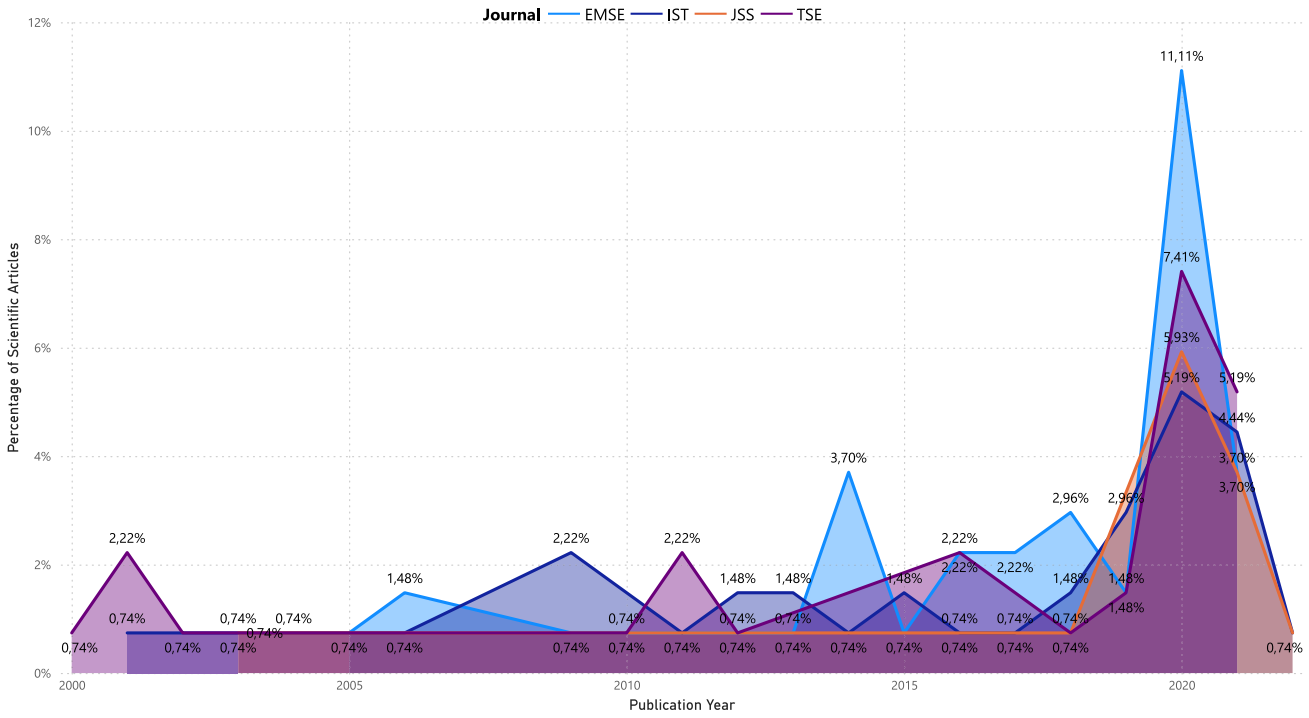

Fig. 4. Trend of journal studies with a defined research artifact from 2000 to 2022 in EMSE, IST, JSS, TSE.

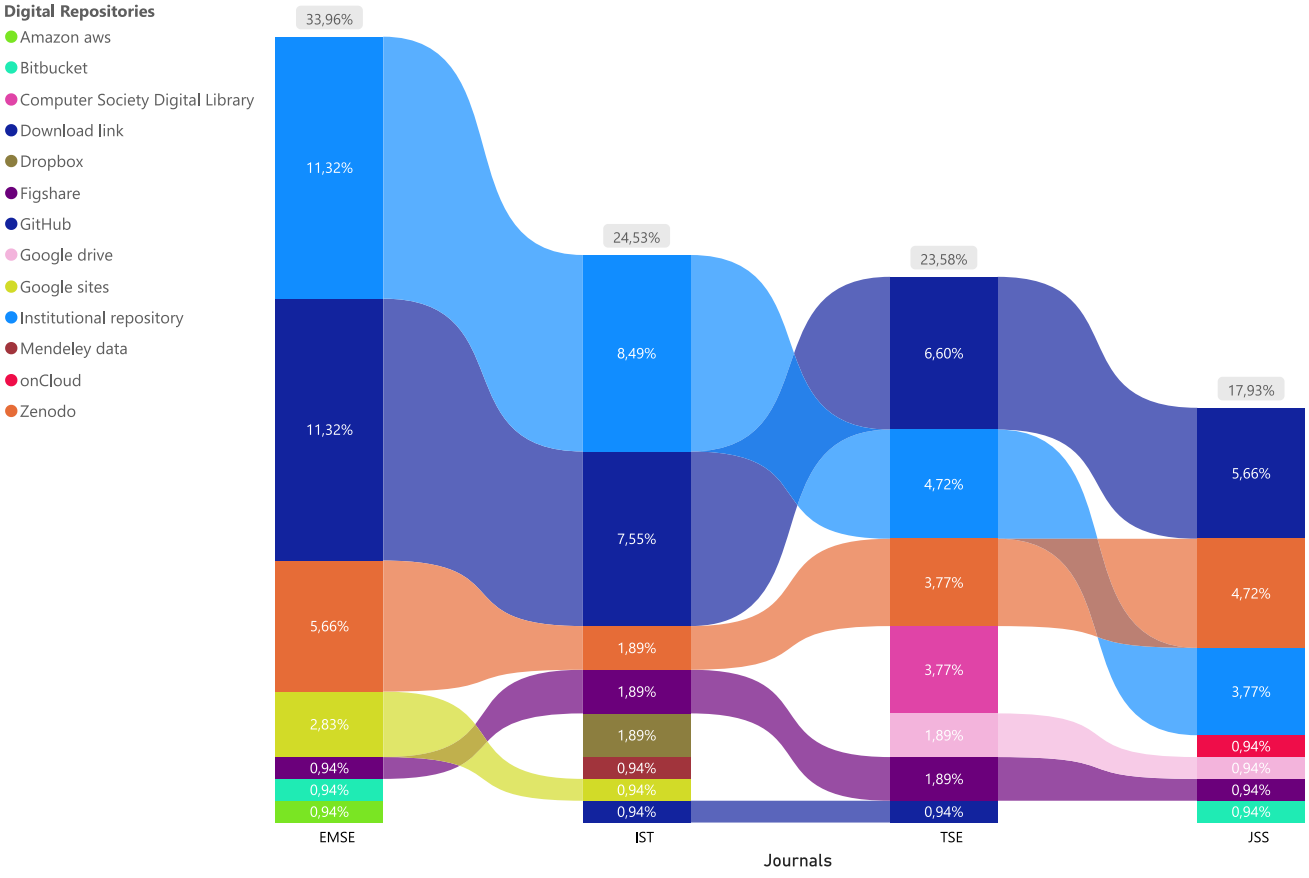

Fig. 5. Most used digital repositories by journals under study.

Table 2
Guidelines to research artifacts in journals under study.

Forum	Description	Prescriptions
EMSE	Works with badges and grants awards called *Open Science Badge*	Web site (EMSE journal, 2023; Springer, 2023)
TSE	Web pages with OS references initiative	Web site (Open Science IEEE Computer Society, 2020)
IST,JSS	Web pages with OS references initiative	Web site (IST Journal, 2023; Elsevier, 2023)

research artifacts very different (Solari, 2011). It can range from very light content such as a summary of the experiment to giving detailed information and instructions on the experimental process (Vegas et al., 2006).

4.4. Digital repositories used to make research artifacts available ($RQ_{1.4}$)

Fig. 5 illustrates the distribution of repositories utilized by authors for disseminating their respective research artifacts. Each bar in the diagram represents a journal, with the total percentage of artifacts showcased atop the corresponding bar. Through the analysis, 13 unique digital repositories were identified, wherein the 106 selected research artifacts were published. Fig. 5 shows these digital repositories alongside their respective usage rates per repository and journal.

Specifically, 33 research artifacts (31,13%) were published in *GitHub*, 30 research artifacts (28,3%) in institutional repositories, and 17 research artifacts (16,04%) in *Zenodo*.

EMSE is the journal where the highest proportion of the journal's papers are uploaded (33.96%), with a percentage of 11.32% in both *GitHub* and *institutional repositories*. Followed by IST with 8,49% in *institutional repositories* and 7,55% in *GitHub*. Among the specific repositories that have emerged in recent years for OS, *Zenodo* is the most used with 16,04%, followed by *Figshare* with 5,66%.

GitHub serves as the predominant digital repository used extensively, accommodating various file formats and providing configuration control. Nevertheless, alternative platforms like *Zenodo* offer the capability to present research artifact content hierarchically by providing a .zip file, as can be seen in Bernárdez et al. (2021). Additionally, Zenodo (OpenAIRE, 2023), as well as Figshare (Figshare, 2023) and OSF (OSF, 2023) among others, assign a DOI (Digital Object Identifier) as a unique identifier for the research artifact and ensure transparent management of multiple versions of the research artifact. Currently, *Zenodo* has more than 10,000 communities created by users, one of them being Empirical SE (Zenodo, 2022). This community is specifically for sharing data, code, experimental material, and papers about empirical knowledge in SE.

The increase of digital repositories currently available on the Internet facilitates the structuring and publication of research artifacts, specifically, the encouragement of the scientific community for the adoption of OS.

4.5. Artifact alignment with ACM badges ($RQ_{2.1}$)

We applied the ACM (2020) badging policies and the (ICSE, 2021) guidelines to identify the main defects that research artifacts suffer from, such as missing relevant information or lack of structure.

Table 3 presents a ranking with ten more common problems identified in the 106 public research artifacts. The second column presents the description of each problem detected and the third column shows the percentage of research artifacts that present such problem. As can be seen in Table 3, *datasets* and the *data analysis scripts* are almost always shared in research artifacts i.e. the percentage of research artifacts presenting these problems is the lowest, while the experimental or protocol design are missing in more than the 88% of the research artifacts (authors usually include these components in a section of the article instead of the research artifact).

The problems shown in Table 3 are closely related to some of the quality characteristics of the ACM badging policy. Firstly, considering that the badge *available* is awarded for research artifact that (i) is shared on the Internet using a DOI and (ii) contains relevant artifacts, the related problems (identified in the first column of Table 3) are mainly *DOI*, *Datasets* and *Material*. Secondly, according to the quality characteristics of *functional* badge such as documented and exercisable, the related problems are mainly *Design*, *Description*, *Protocol*, and *Analysis*. Finally, considering that the *reusable* badge adds two quality characteristics (structured and well-documented) to the *functional* badge, the more related problems are *Websites*, *Structure*, and *Software*, apart from those in the *functional* badge.

Table 3
Top ten problems detected in the journal research artifacts concerning to ACM badging policy.

Problem	Problem description	% of research artifacts with the problem
Design	The description of the experimental design is missing in the research artifact.	91,5%
Protocol	The description of the experimental protocol is missing in the research artifact.	88,8%
Structure	Files are not organized by category or type.	75,5%
DOI	The research artifact is available on the Internet but without a DOI.	62,3%
Material	Questionnaires, surveys, or training material are missing in the research artifact.	59,4%
Description	Explanations about the composition and structure of the research artifact are missing.	47,2%
Software	The research artifact contains any software without guidelines or rationale.	46,2%
Websites	The research artifact is available on the Internet and distributed on several websites.	22,6%
Analysis	The data analysis scripts are missing in the research artifact.	20,8%
Datasets	The datasets are missing in the research artifact or they are not accessible.	6,6%

4.6. Completeness of artifacts: Solari's module ($RQ_{2.2}$)

In this question, we have studied how complete research artifacts were with respect to the proposal of Solari et al. (2018), organized in seven modules, as summarized previously in Section 2.3.2. Each module contains artifacts that have been carefully developed so that researchers can use them to understand the experiment and replicate it.

Analyzing the 106 research artifacts selected, their content was categorized into *complete* and *not complete*. As seen in Fig. 6, within the set of Solari et al. (2018) modules, three modules are the most frequently made available to the community: *experiment, training* and *introduction*. 100% of the research artifacts present the *experiment* module, although only 35,85% have the complete level.

The content associated with the module *training* is only presented by 49,15% of the research artifacts. Although only 17,8% have the complete level. The rest of the research artifacts (50,85%) do not present any content of this module. Besides, the *theory* module is rarely reported because the authors report very few conceptual models, diagrams, bibliographic references, descriptions of the techniques, and formulas used in the experiment. Nevertheless, the *aggregation, replication*, and *evolution* modules only appear in research artifacts belonging

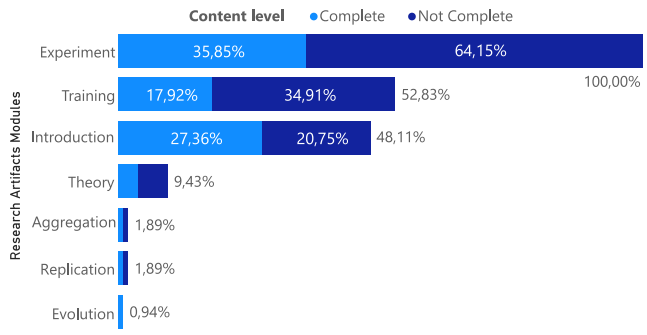

Fig. 6. Research artifacts from journals classified in modules proposed by Solari et al. (2018).

to replications or a *family of experiments*, not an original experiment. Since most of the research artifacts analyzed report material from original experiments, the three modules mentioned are less common.

The most common content uploaded by the authors, related to the *experiment* module, is the operational material used in the experiment, e.g., forms, data collection, test cases, questionnaires, and statistical analysis results. For example, those developed by Siegmund et al. (2014) and Prechelt (2011). Furthermore, in the *training* module, there is a percentage of works that do not publish anything related to training due to two factors. The first factor is that the experiments do not need to execute training tasks for the subjects. Secondly, experimenters do not upload related artifacts even though they perform training tasks. In the case of publishing the *training* module, the most common is to share theoretical presentations and training guides, registration of preliminary exercises before conducting the experiment, and registration of feedback sessions. Examples of *training* module packages are those developed by Ricca et al. (2018) and Fakhoury et al. (2020).

Like the preceding module, the *introduction* module presents a percentage of works that do not publish anything related. In the case of publishing the *introduction* module, the most usual is that experimenters upload a record of the research artifact organizational structure, communication instruments between experimenters, and contact information of the original experiment and replication authors. Examples of *introduction* module packages are those developed by Matias Urbieta et al. (2020) and Fernández-Sáez et al. (2016). Only one article by Basili et al. (2006) was found whose research artifact completes all the modules.

5. HOE research artifact structure

In this section, we present not only the basic structure of HOE research artifacts (Table 4) but also the compliance of our proposal according to ACM (2020) and Solari et al. (2018) proposals (Tables 5 and 6).

As can be seen in Table 4, the proposed research artifacts structure comprises a list of items classified into the following sections, *Article*, *Experimental material*, *Datasets*, *System*, *Scripts*, *Dictionary* which contains metadata about research artifact, and finally, *Reproducibility and Replicability Information*, is an optional section that provides evidence showing whether the experiment has been replicated or reproduced by a subsequent study. Note that in Table 4 the artifacts are organized into sections, allowing to achieve the *well-structured* quality characteristic of the *reusable* badge. The other characteristic of that badge, *carefully documented*, is achieved by filling in the elements of the *Dictionary*.

In Table 5, a checkmark (✓) represents that an item is recommended to obtain the *functional, reusable*, and *available* ACM badges, as indicated by the examination of the sources of our survey (Abrahão and Mendez, 2021; Solari and Vegas, 2006; Solari et al., 2018; Basili et al., 1996), while an asterisk (*) represents that an item that we consider as mandatory, as explicitly stated in the recommendations provided by ACM (2020).

In Table 6, the optionality character of the items in section *Reproducibility and Replicability Information* is presented, for the *reproduced* and *replicated* ACM badges.

Finally, in addition to information about ACM badges, Tables 5 and 6 show which level of maturity each item helps to achieve, according to Solari and Vegas (2006). Our goal is to help experimenters develop an organized and structured research artifact with an acceptable level of completeness to award one or more ACM badges.

5.1. Validation of the research artifact structure

We compared our structure proposal with respect to 12 research artifacts that applied for ACM badges in relevant conference AE tracks. Before starting the comparison, we checked that none of the 12 research artifacts were among those selected for the survey presented in the current work. The validation process involved the following steps: (i) reviewing the mandatory items to confirm their relevance, (ii) checking that recommended items are less demanded than mandatory ones and (iii) verifying that research artifacts without recognition in the corresponding AE track lacked both mandatory and recommended items. Table 7 summarizes the result of the validation analysis, showing that of the 12 research artifacts that applied for badges, 7 earned the *available* badge, 6 earned the *reusable* badge, and 4 did not receive any badge at all.

In addition, we validated the proposal with research artifacts from two own experiments (Bernárdez et al., 2023; Uyaguari et al., 2023). In this validation, we verified that the *experimental material* section of the (Bernárdez et al., 2023) experiment is strong. We also observed that the *scripts* were made in a Jupyter notebook, but there were no guides to run them. In the (Uyaguari et al., 2023) research artifact, we observed that artifacts were missing in the *scripts* section, but they had the mandatory *research artifact dictionary* items. Finally, the *system* section is missing in both experiments because the experiments do not use software.

5.1.1. Validation of available badge structure

As can be seen in the second column of Table 7, the 6 mandatory items for the *available* badge in our proposal are *raw data, script description, analysis scripts, research artifact description, DOI* and *public repository*, displayed in boldface and marked with the corresponding badge icon.

Of the 7 research artifacts with the available badge in Table 7, 4 contain all the 6 mandatory items, i.e., 100% coverage. From the remaining 3 research artifacts (A, B, G), A and B lack a *DOI* and, moreover, in B the item *research artifact description* is missing.

Nevertheless, in our proposal we have considered the *research artifact description* as mandatory because, apart from appearing in most of the research artifacts awarded with the *available* badge, it describes each of the items, providing important information to the experimenters about the context of the experiment and the items used in each phase of the experimental process. In addition, at other conferences such as ICSE, there is a strong emphasis on describing the items for better research artifact comprehensibility.

Consequently in our proposal, A, B, and G research artifacts show respectively a coverage of 83.3%, 66.6%, and 55.5% of the mandatory items for the *available* badge.

Based on the research artifacts we have examined, it appears that the items we initially deemed mandatory hold significant importance. However, note that it is still possible to obtain the *available* badge without including all of these items.

On the other hand, we can see that research artifacts I, J, K, and L, where *public repository* is registered, were not granted the available badge. These research artifacts lack some required items such

Table 4
Research artifact structure proposal for HOE in SE.

Section & Items	Description
Article	
Scientific article	Article file.
Design description	Experimental design specification file.
Context description	File describing the environment in which the experiment was carried out and its features.
Author registration	Text with the author's name, e-mail, address, and affiliation.
Experimental material	
Participant instructions	Forms, statements, or any file for developing the experimental tasks.
Participant demographic surveys	File of forms addressed to participants to collect information on their characteristics.
Training	File with the material used during the training process or learning sessions.
Measuring instruments	File with instruments used to provide reliable measurements such as data analytics tools.
Conceptual models	File with diagrams that explain the theory underlying the experiment.
Datasets	
Raw data	File that include the records of the experiment data before manipulation.
Curated data	File that include the records of the experiment data after manipulation.
System	
Software description	File describing the functionalities and specifications of the software.
Successful software execution	Result file with the result of the execution of the software.
Software technical requirement	Description file of software technical constraints, such as the development environment.
Hardware technical requirement	Description file of hardware technical constraints, such as screen resolution.
Scripts	
Script description	File describing the source code of the data analysis scripts.
Analysis scripts	File describing data analysis that may be in a conventional format or employ virtualization or containerization, as recommended (Timperley et al., 2021).
Successful script execution	A log file generated by the author with the correct script execution.
Research artifact dictionary	
Inventory	File with a list of all items.
Description	File describing the items used in each phase of the experimental process.
Zipped file	Compressed file that groups items in a structured way.
Virtual environments	File description of the computer-simulated digital spaces used in the experiment.
Distribution rights	File with the legal permits granted for the distribution of the research artifact.
Installation guideline	File with step-by-step instructions for the software installation.
License available	Text that includes the software license availability.
Type (Simple and/or installation)	Text with the selected type. Simple consists only of documentary elements such as data files or analysis scripts. The installation type consists of a tool or software, either the software under study or the software used by the participants in the experimental tasks.
Justification for claiming the badge	Text with the reasons why the authors claim one or more badges.
Configuration guideline	File with step-by-step instructions for the software configuration.
DOI	URL with the DOI of the research artifact which prevents the appearance of broken links.
Accessible data	Its value is *Yes* when the data is accessible for reading and *No* otherwise.
Manipulable data	Its value is *Yes* when the data is accessible for writing and *No* otherwise.
Public repository	URL providing the link where the research artifact is available (e.g. an institutional repository or Github).
Confidential access	Its value is *Yes* when the access to the research artifact is protected by a password and *No* otherwise.
Download guideline	File with step-by-step instructions for downloading the research artifact.
Execution guideline	File with step-by-step instructions for the software execution.
Decision-making	File that summarizes motivation and rationale that led the authors to conduct the experiment.
Reproduction and Replication Information	
Reproduction or Replication article	Article file in which the subsequent study (replication or reproduction) is reported.
Reproduction or Replication authors	Text with data from the subsequent study authors.
Summary	Text with a brief description of how the subsequent study was carried out.
Substantial proofs	Its value is YES when there is evidence that the subsequent study has been carried out, e.g. the existence of a replication or reproduction research artifact.
Related work	File comprising the bibliography used to design and perform the subsequent study.
Ethic edge	Its value is YES when the subsequent study refrains from making negative criticisms of prior works.
Tolerance margin	File describing the results of the subsequent study, which should be close to the results of the original experiment within a tolerance considered acceptable for experiments of the given type.

as *accessible data, manipulable data* or *curated data*. That is, the *public repository* alone is not sufficient to achieve the *available* badge.

Finally, note that research artifact E did not obtain the *available* badge but shows great coverage compared to the rest of the awarded research artifacts. A plausible explanation is that research artifact E lacks the *research artifact description* and *instruction guidelines* items, which are considered essential in the ICSE AE track guidelines (Abrahão and Mendez, 2021).

5.1.2. Validation of reusable badge structure

Regarding the 8 mandatory items for the *reusable* badge in our proposal (*raw data, script description, analysis scripts, inventory, research artifact description, accessible data, download instruction guideline* and *execution instruction guideline*), 4 out of the 6 research artifacts with the *reusable* badge in Table 7 present a 100% coverage. However, in the remaining 2 research artifacts, B and E, the *research artifact description, download instruction guideline* and *execution instruction guideline* are missing, presenting a 62% coverage.

It can also be seen that research artifacts G and H have not obtained the *reusable* badge because they lack mandatory items such as *download instruction guideline*, *execution instruction guideline*, and *inventory*. Note that the latter is included in all research artifacts that have received the *reusable* badge. We consider that these 3 missing items contribute to making the research artifact *well-documented*, as it is one of the quality characteristics that ACM guidelines requires for the *reusable* badge.

In summary, after an in-depth study of the literature and a review of research artifacts presented in AE tracks, it should be noted that conferences do not always require complete fulfillment of the core set in our proposal to grant the *reusable* badge.

Table 5
Mapping of the research artifact structure proposal to the *functional*, *reusable* and *available* ACM badges and Solari's maturity levels.

Section	Items	Functional	Reusable	Available	Maturity levels
Article	Scientific article	✓	✓	✓	Article
	Design description	✓	✓	–	Descriptive
	Context description	✓	✓	–	Descriptive
	Author registration	–	✓	✓	Descriptive
Experimental material	Participant instructions	✓	✓	✓	Operational
	Participant demographic surveys	✓	✓	–	Operational
	Training	✓	✓	✓	Operational
	Measuring instruments	✓	✓	–	Operational
	Conceptual models	✓	✓	–	Operational
Datasets	Raw data	*	*	*	Operational
	Curated data	✓	✓	–	Operational
System	Software description	✓	✓	–	Operational
	Successful software execution	✓	✓	–	Operational
	Software technical requirement	–	✓	–	Operational
	Hardware technical requirement	–	✓	–	Operational
Scripts	Script description	*	*	*	Procedural
	Analysis scripts	*	*	*	Procedural
	Successful scripts execution	✓	✓	–	Procedural
Research artifact dictionary	Inventory	✓	*	–	Descriptive
	Description	*	*	*	Descriptive
	Zipped file	–	✓	–	Descriptive
	Virtual environments	–	✓	–	Descriptive
	Distribution rights	–	✓	✓	Descriptive
	Installation guideline	–	✓	–	Procedural
	Licence available	–	✓	✓	Descriptive
	Type (Simple and/or installation)	–	✓	–	Descriptive
	Justification for claiming the badge	–	✓	✓	Descriptive
	Configuration and installation	–	✓	–	Descriptive
	DOI	–	–	*	Descriptive
	Accessible data	*	*	–	Operational
	Manipulable data	✓	✓	–	Operational
	Public repository	–	–	*	Descriptive
	Confidential access	–	–	✓	Descriptive
	Download instruction guideline	*	*	–	Procedural
	Execution instruction guideline	*	*	–	Procedural
	Decision-making	✓	✓	–	Decision-making
	Total Items	22	33	14	

Table 6
Mapping of the research artifact structure proposal to the *reproduced* and *replicated* ACM badges and Solari's maturity levels.

Section	Item	Reproduced	Replicated	Maturity levels
Article	Scientific article	✓	✓	Article
	Reproduction article	✓	–	Article
	Reproduction authors	✓	–	Descriptive
	Author registration	✓	✓	Descriptive
	Replication article	–	✓	Article
	Replication authors	–	✓	Descriptive
Research artifact dictionary	Summary on reproduction	✓	–	Descriptive
	Substantial test on reproduction	✓	–	Descriptive
	Related work on reproduction	✓	–	Descriptive
	Ethic edge on reproduction	✓	–	Descriptive
	Tolerance margin on reproduction	✓	–	Descriptive
	Summary on replication	–	✓	Descriptive
	Substantial tests on replication	–	✓	Descriptive
	Related work on replication	–	✓	Descriptive
	Ethic edge on replication	–	✓	Descriptive
	Tolerance margin on replication	–	✓	Descriptive
	Total items	9	9	

5.1.3. Validation of research artifacts without badge

Focusing on the 4 research artifacts (I, J, K, and L) in Table 7 that have not received any badges, the main lack with respect to our proposal is the absence of a number of mandatory items. Specifically, (i) the I research artifact does not include *datasets, DOI, accessible data, download instruction guideline*, or *execution instruction guideline*, (ii) the J and K research artifacts do not include *scripts section, DOI, download instruction guideline*, or *execution instruction guideline*, and (iii) the L research artifact has minimal *datasets, scripts* or *research artifact dictionary* which are all mandatory items in our proposal.

Some of the aforementioned research artifacts also lack the items of the *system* section. We consider that this section depends on the type of experiment since if the experiment does not include software development, this section is not required. As a result, its items are considered as recommended (✓) in our proposal.

Table 7
Analysis of the proposed research artifact structure with respect to research artifacts in conference AE tracks (reusable, available).

Section	Items	ASE		ICSE			ESEC/FSE						RE
		(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)
Article	Scientific article	✓	✓	–	–	–	–	–	–	–	–	–	–
	Design description	–	✓	–	–	–	–	–	–	–	–	–	–
	Context description	–	✓	–	–	–	–	–	–	✓	–	–	–
	Author registration	✓	✓	–	✓	✓	-	–	✓	✓	–	–	✓
Experimental material	Participant instructions	–	–	–	–	–	–	–	–	✓	✓	–	✓
	Participant demographic surveys	–	–	–	–	–	–	–	–	✓	-	–	✓
	Training	✓	–	–	–	✓	–	–	–	✓	✓	–	–
	Measuring instruments	–	–	–	–	–	–	–	–	–	–	–	–
	Conceptual models	–	✓	–	–	–	–	–	–	–	–	–	–
Datasets	**Raw data**	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	✓
	Curated data	✓	✓	–	✓	–	✓	✓	✓	–	✓	–	–
System	Software description	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	–	–
	Successful software execution	✓	✓	✓	✓	✓	✓	✓	✓	–	–	–	–
	Software technical requirement	✓	✓	✓	✓	✓	✓	–	✓	–	–	–	–
	Hardware technical requirement	✓	✓	✓	✓	✓	✓	–	–	–	–	–	–
Scripts	**Scripts description**	✓	✓	✓	✓	✓	✓	–	✓	✓	–	–	–
	Analysis scripts	✓	✓	✓	✓	✓	✓	–	✓	✓	–	–	✓
	Successful scripts execution	✓	✓	✓	✓	✓	✓	–	✓	✓	–	–	–
Research artifact dictionary	**Inventory**	✓	✓	✓	✓	✓	✓	–	–	✓	–	–	–
	Description	✓	–	✓	✓	–	✓	–	✓	✓	✓	✓	–
	Zipped file	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓	✓
	Virtual environments	–	✓	–	–	✓	–	–	–	–	–	–	–
	Distribution rights	✓	–	✓	✓	✓	✓	–	✓	–	–	–	–
	Installation guideline	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–
	Licence available	✓	✓	–	✓	✓	✓	✓	✓	–	–	–	–
	Type (Simple/installation)	–	–	–	–	–	–	–	–	–	–	–	–
	Justification for claiming the badge	–	–	–	–	–	–	–	–	–	–	–	–
	Configuration and installation	✓	–	✓	✓	–	✓	–	–	–	–	–	–
	DOI	–	–	✓	✓	✓	✓	✓	✓	–	-	–	✓
	Accessible data	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	–	–
	Manipulable data	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	–	–
	Public repository	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Confidential access	–	–	–	–	–	–	–	–	–	–	–	–
	Download instruction guideline	✓	–	✓	✓	–	✓	–	–	–	–	–	–
	Execution instruction guideline	✓	–	✓	✓	–	✓	–	–	–	–	–	–
	Decision–making	–	–	–	–	–	–	–	–	–	–	–	–
	Total Items	24	22	20	23	20	22	10	17	12	10	4	8

(A) Paltenghi and Pradel (2021); (B) Shi et al. (2021b); (C) Braz et al. (2021); (D) Peitek et al. (2021); (E) Cooper et al. (2021); (F) Shen et al. (2021); (G) Jiang et al. (2021); (H) Shi et al. (2021a); (I) Endres et al. (2021); (J) Liu et al. (2021); (K) Wang et al. (2021); (L) Gralha et al. (2021).

The 4 non-awarded packages have a reduced content compared to the awarded ones, thus we note that they present very few items in the *research artifact dictionary* section, which is very important for the *reusable* badge. As for the reasons why they have not received the *available* badge, despite having *public repository* (which is essential), almost none of them present *DOI* and have very few items in *scripts* section, when the 4 research artifacts have statistical analysis and software developed.

In summary, after applying our structure proposal to 12 research artifacts that applied for ACM badges in relevant conference AE tracks, the results confirm that most of the items considered mandatory in our proposal are relevant to achieving the corresponding badge. Moreover,

each item encompassed within the analyzed research artifacts corresponds to a specific section within our proposal. Finally, the coverage that the structure proposal has in those research artifacts is in mean 85.6% for the *available* badge and 85.7% for the *reusable* badge.

6. Threats to validity

Like any empirical study, the presented research has limitations in scope and threats to validity (Wohlin et al., 2012). We have considered *external validity* threats, which affect the generalizability of the results, and *internal validity* threats, which deal with the examination of causal relationships (Siegmund et al., 2015). Moreover, we considered potential threats to the validity of secondary studies (Ampatzoglou et al., 2019) in the conduct of the survey and in the validation of the proposed research artifact structure following the DSR process (Peffers et al., 2007).

The search for experiments focused on four relevant journals; this limitation constitutes a threat to the *external validity* of the survey. Nevertheless, this search allowed us to access a considerable number of experiments (106 primary studies) that were probably of good quality in their experimental design and publication (due to the peer validation process and the journals considered). On the other hand, this search in journals covers the last 20 years. That is, it is considered a long period of time that allows us to have a fairly broad version of the quality of the research artifacts. The papers considered start a few years after (Basili et al., 1996, 1999) and, therefore, it is to be expected that this proposal has influenced the historical evolution of research artifacts in the selected experiments.

This study focuses on a subtype of experiments (HOE). We have tried to have the initial search string return as many articles as possible and then processed a large number of sources to choose HOE. The inclusion criterion has been applied manually (from the initial 1426 articles, we were left with 267 that dealt with HOE). Although the manual selection is prone to error, entering the term HOE in the initial search string would have decreased the recall of the search. It is unlikely that experiments would self-classify using the term 'human', so the study would have been inappropriately excluded even if it was an HOE.

There are threats to *internal validity* that arise from the process of analyzing the research artifacts of the selected experiments. In most cases, the research artifacts were available in digital repositories, so the review process focused on the content that the authors had uploaded to the repository. However, in other cases, the research artifacts were only described in the article and distributed in a non-homogeneous manner. On the other hand, the expert judgment applied to categorize research artifacts according to different guidelines establishes a bias in the research. To mitigate this threat to validity, a prior definition of the categorization criteria has been made. The initial categorization made by the first author has been iteratively revised by the second and third co-authors.

As a result of the DSR process, the proposed research artifact structure has biases based on the specific research context in which it was designed and validated. The proposal takes into consideration the main OS guidelines of the ESE community while focusing on HOE. The proposal was validated using two complementary approaches. First, the proposal was benchmarked by analyzing the content of research artifacts in 12 contemporary experiments that received different levels of ACM badges. On the other hand, a direct empirical analysis was performed by applying the proposal in two experiments (Bernárdez et al., 2023; Uyaguari et al., 2023). This could have caused limitations because the proposal's authors are the same as we made the research artifacts of both experiments, as defined in researcher bias (Chenail, 2014). Nevertheless, when a DSR artifact was proposed, using the known context of our own experiments allowed us to perform an initial field evaluation.

7. Conclusions and future work

This article briefly describes the main proposals that have emerged to advance the OS movement in the ESE community. It then analyzes how these initiatives have been received by the main journals and conferences in the field, and what actions they have taken to encourage authors to follow OS initiatives, claiming the interest in developing quality research artifacts of their empirical studies and making them public to the community. In order to carry out this analysis, the survey was applied to 106 research artifacts, concluding that, (i) 39,70% (106/267) of journal studies that reported a HOE between 2000 and 2022 make completely accessible their research artifacts. The main problems with accessibility are due to broken URLs or repositories requiring authorization. Furthermore, authors that publish in journals without pages limit, decide to describe all the artifacts in the manuscript itself, as occurs in EMSE journal ($RQ_{1.1}$), (ii) with respect to journals, we find an average of 5 research artifacts per year, and in 2020 there is a considerable increase to 35 research artifacts. We conclude that the trend in the publication of research artifacts is growing ($RQ_{1.2}$); (iii) journals have published on their websites important calls for OS, and conferences, based on the (ACM, 2020) initiative, have carried out detailed guidelines to develop good research artifacts. In other words, the ESE community is making a big effort in standardization, review, and digital publishing related to OS. Nevertheless, the research artifacts reviewed during the survey do not usually follow these guidelines nor other relevant proposals such as those by Basili and Selby (1991) and Solari and Vegas (2006), so the resulting research artifacts present some lack of completeness ($RQ_{1.3}$ and $RQ_{2.1}$); (iv) of the 13 digital repositories identified where the authors upload their research artifacts, the most commonly used in the SE community are *GitHub*, institutional repositories, and *Zenodo* ($RQ_{1.4}$); (v) The most common items in the research artifacts are instructions about experimental settings, operating material (that included training of participants) and experimental tools.

In order to internally validate the structure proposal, we have applied the proposal to 12 research artifacts, which have been awarded at conferences with ACM badges. In future work, we aim to validate the proposed approach through external validation by applying it to a series of research artifacts (both HOE and computational experiments) in order to evaluate its goodness. Additionally, we intend to develop a method for constructing research artifacts during the experimental process itself, taking advantage of the availability of all necessary materials and enabling experimenters to create the packages with greater precision and efficiency. Furthermore, we will develop a software tool that not only supports this method but also assists experimenters in building an artifact structured according to the requirements of the ACM badges.

Research artifact

The research artifact of this work is available at the Zenodo repository (Guevara-Vega et al., 2024).

CRediT authorship contribution statement

Cathy Guevara-Vega: Conceptualization, Supervision, Project administration, Methodology, Formal analysis, Investigation, Funding acquisition, Data curation. **Beatriz Bernárdez:** Conceptualization, Supervision, Methodology, Formal analysis, Data curation. **Margarita Cruz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Data curation. **Amador Durán:** Writing – original draft, Visualization, Resources, Project administration. **Antonio Ruiz-Cortés:** Conceptualization, Supervision, Resources, Funding acquisition. **Martin Solari:** Writing – review & editing, Validation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We would like to thank the anonymous reviewers who have made very useful comments to improve the work.

This work has been partially supported by the Universidad Técnica del Norte–Ecuador, and by grants PID2021-126227NB-C21 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE, TED2021-131023B-C21 and C22 funded by MCIN/AEI/10.13039/501100011033.

References

Abrahão, S., Insfran, E., Guevara, L., González, F., Fernández, M., Cano, C., Oliveira, R., 2019. Assessing the effectiveness of goal-oriented modeling languages: A family of experiments. Inf. Softw. Technol. 116, http://dx.doi.org/10.1016/j.infsof.2019.08.003.

Abrahão, S., Mendez, D., 2021. ICSE'2021 artifact evaluation track-submission and reviewing guidelines. https://bit.ly/3yyfh32.

ACM, 2020. Artifact review and badging-current. https://bit.ly/3oQqJBO.

Almqvist, J., 2006. Replication of Controlled Experiments in Empirical Software Engineering-A Survey (Ph.D. thesis). Lund University, pp. 53–80, URL: https://bit.ly/31Y5OUz.

Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M., Chatzigeorgiou, A., 2019. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. Inf. Softw. Technol. 106, 201–230.

Basili, V., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sorumgard, S., Zelkowitz, M., 1996. Packaging researcher experience to assist replication of experiments. In: Proc. of the ISERN Meeting 1996.

Basili, V., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sørumgård, S., Zelkowitz, M., 2006. Lab package for the empirical investigation of perspective-based reading. URL: bit.ly/475MOR6.

Basili, V., Selby, R., 1991. Paradigms for experimentation and empirical studies in software engineering. Reliab. Eng. Syst. Saf. (32), 171–191. http://dx.doi.org/10.1016/0951-8320(91)90053-A.

Basili, V., Shull, F., Lanubile, F., 1999. Building knowledge through families of experiments. IEEE Trans. Softw. Eng. 25 (4), 456–473.

Bernárdez, B., Panach, I., Parejo, J., Durán, A., Juristo, N., Ruiz-Cortés, A., 2021. An empirical study to evaluate the impact of mindfulness at a software company (laboratory package). http://dx.doi.org/10.5281/zenodo.5723957.

Bernárdez, B., Panach, I., Parejo, J., Durán, A., Juristo, N., Ruiz-Cortés, A., 2023. An empirical study to evaluate the impact of mindfulness on helpdesk employees. Sci. Comput. Program. http://dx.doi.org/10.1016/j.scico.2023.102977.

Blohowiak, B., Cohoon, J., de Wit, L., Eich, E., Farach, F., Hasselman, F., Holcombe, A., Humphreys, M., Lewis, M., Nosek, B., Peirce, J., Spies, J., Seto, C., Bowman, S., Green, D., Nilsonne, G., Grahe, J., Wykstra, S., Hofelich, A., Andrew, S., Giner-Sorolla, R., Parker, T., Forstmeier, W., Nakagawa, S., Kidwell, M., Mellor, D., DeHaven, A., Riss, C., Lowrey, O., 2022. Badges to acknowledge open practices. https://osf.io/tvyxz/.

Braz, L., Fregnan, E., Çalikli, G., Bacchelli, A., 2021. Data and materials for: Why don't developers detect improper input validation?'; DROP TABLE papers; –. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings. ICSE-Companion, pp. 203–204. http://dx.doi.org/10.1109/ICSE-Companion52605.2021.00090.

Brooks, A., Daly, J., Miller, J., Roper, M., Wood, M., 1994. Replication's role in experimental computer science. EfoCS–5–94 (RR/94/171).

CERN, 2023. European organization for nuclear research. https://home.cern/.

Chenail, R., 2014. Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research. Qual. Rep. http://dx.doi.org/10.46743/2160-3715/2011.1051.

Conradi, R., Basili, V., Carver, J., Shull, F., Travassos, G., 2001. A Pragmatic Documents Standard for an Experience Library: Roles, Documen, Contents and Structure. Technical Report CSTR-4235, University of Maryland, pp. 1–50, URL: http://go.umd.edu/3YaJKPt.

Cooper, N., Bernal-Cárdenas, C., Chaparro, O., Moran, K., Poshyvanyk, D., 2021. A replication package for it takes two to tango: Combining visual and textual information for detecting duplicate video-based bug reports. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings. ICSE-Companion, pp. 160–161. http://dx.doi.org/10.1109/ICSE-Companion52605.2021.00067.

Cruz, M., Bernárdez, B., Durán, A., Galindo, J.A., Ruiz-Cortés, A., 2019. Replication of studies in empirical software engineering: A systematic mapping study, from 2013 to 2018. IEEE Access 8, 26773–26791. http://dx.doi.org/10.1109/ACCESS.2019.2952191.

Dasgupta, P., Paul, D., 1994. Toward a new economics of science. Res. Policy 23 (5), 487–521. http://dx.doi.org/10.1016/0048-7333(94)01002-1.

De Magalhães, V., Da Silva, F., 2013. Towards a taxonomy of replications in empirical software engineering research: A research proposal. In: Proceedings - 2013 3rd International Workshop on Replication in Empirical Software Engineering Research. RESER 2013, pp. 50–55. http://dx.doi.org/10.1109/RESER.2013.10.

Elsevier, 2023. Open data report: The researcher perspective. https://bit.ly/3Y8khpK.

EMSE, Empirical software engineering journal. https://bit.ly/3QeQLwx.

EMSE journal, 2023. The open science initiative of empirical software engineering. https://bit.ly/3zgdeiZ.

Endres, M., Fansher, M., Shah, P., Weimer, W., 2021. To read or to rotate? Comparing the effects of technical reading training and spatial skills training on novice programming ability. In: To Read Or To Rotate? Comparing the Effects of Technical Reading Training and Spatial Skills Training on Novice Programming Ability. In: ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA, pp. 754–766. http://dx.doi.org/10.1145/3468264.3468583.

Fakhoury, S., Roy, D., Ma, Y., Arnaoudova, V., Adesope, O., 2020. Measuring the impact of lexical and structural inconsistencies on developers' cognitive load during bug localization. Empir. Softw. Eng. 25, http://dx.doi.org/10.1007/s10664-019-09751-4.

Fernández-Sáez, A., Genero, M., Caivano, D., Chaudron, M., 2016. Does the level of detail of UML diagrams affect the maintainability of source code? A family of experiments. Empir. Softw. Eng. 21 (1), 212–259, URL: https://doi.org/10.1007/s10664-014-9354-4.

Ferro, N., Kelly, D., 2018. SIGIR initiative to implement ACM artifact review and badging. In: ACM SIGIR Forum. Vol. 52, ACM New York, NY, USA, pp. 4–10.

Figshare, 2023. Figshare repository. https://figshare.com/.

FORCE11, 2021. Guiding principles for findable, accessible, interoperable and reusable data publishing. https://bit.ly/3vmFPBT.

Gómez, O., Juristo, N., Vegas, S., 2010. Replication, reproduction and re-analysis: Three ways for verifying experimental findings. In: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research (RESER 2010), Cape Town, South Africa.

Gralha, C., Pereira, R., Goulão, M., Araujo, J., 2021. On the impact of using different templates on creating and understanding user stories. In: 2021 IEEE 29th International Requirements Engineering Conference. RE, pp. 209–220. http://dx.doi.org/10.1109/RE51729.2021.00026.

Graziotin, D., 2021. Open science policies. https://bit.ly/31UNn39.

Guevara-Vega, C., Bernárdez, B., Cruz, M., Durán, A., Ruiz-Cortés, A., Solari, M., 2024. Supplemental material: Research artifacts for human-oriented experiments in software engineering: An ACM badges–driven structure proposal. https://doi.org/10.5281/zenodo.11222784.

Guevara-Vega, C., Bernárdez, B., Durán, A., Quiña-Mera, J., Cruz, M., Ruiz-Cortés, A., 2021. Empirical strategies in software engineering research: A literature survey. In: II International Conference on Information Systems and Software Technologies. ICI2ST 2021, IEEE Press, Quito, Ecuador, http://dx.doi.org/10.1109/ICI2ST51859.2021.00025.

Hermann, B., Winter, S., Siegmund, J., 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. pp. 469–480.

Heumüller, R., Nielebock, S., Krüger, J., Ortmeier, F., 2020. Publish or perish, but do not forget your software artifacts. Empir. Softw. Eng. 25 (6), 4585–4616. http://dx.doi.org/10.1007/S10664-020-09851-6.

ICSE, 2021. AE-Artifact Evaluation. https://bit.ly/3JxuQeu.

IEEE, 2023. IEEE dataport. https://bit.ly/3Kd4tfO.

IST, Information and software technology journal. https://bit.ly/3H0Zhss.

IST Journal, 2023. Free preprint posting service in SSRN (open-access online pre-print server). https://bit.ly/3RYMgom.

Jiang, Y., Liu, H., Zhang, Y., Niu, N., Zhao, Y., Zhang, L., 2021. Which Abbreviations Should Be Expanded? In: ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA, pp. 578–589. http://dx.doi.org/10.1145/3468264.3468616.

JSS, Journal of systems and software journal. https://bit.ly/3tlGGSy.

JSS Journal, 2023. JSS open science. https://bit.ly/3q4r8E5.

Juristo, N., Gómez, O.S., 2010. Replication of software engineering experiments. In: Empirical Software Engineering and Verification. Springer, pp. 60–88.

Kamsties, E., Lott, C., 1995a. An empirical evaluation of three defect-detection techniques. In: Proceedings of the Fifth European Software Engineering Conference. ESEC.

Kamsties, E., Lott, C., 1995b. An Empirical Evaluation of Three Defect-Detection Techniques. Technical Report ISERN 95–02, Department of Computer Science, University of Kaiserslautern, 67653 Kaiserslautern, Germany.

Liu, M., Peng, X., Marcus, A., Treude, C., Bai, X., Lyu, G., Xie, J., Zhang, X., 2021. Learning-Based Extraction of First-Order Logic Representations of API Directives. In: ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA, pp. 491–502. http://dx.doi.org/10.1145/3468264.3468618.

Matias Urbieta, M., Leandro Antonelli, L., Gustavo Rossi, G., Sampaio do Prado Leite, J., 2020. The impact of using a domain language for an agile requirements management. Inf. Softw. Technol. 127, 106375. http://dx.doi.org/10.1016/j.infsof.2020.106375.

Mendeley Data, 2023. Datasets associated with articles published in journal of systems and software. https://bit.ly/44FEUfj.

Méndez, D., Graziotin, D., Wagner, S., Seibold, H., 2020. Open science in software engineering. Contemp. Empir. Methods Softw. Eng. 477–501. http://dx.doi.org/10.1007/978-3-030-32489-6_17, arXiv:1904.06499.

Méndez, D., Monperrus, M., Feldt, R., Zimmermann, T., 2019. The open science initiative of the empirical software. Empir. Softw. Eng. 1 (24), 1057–1060. http://dx.doi.org/10.1007/s10664-019-09712-x.

Mian, P., Travassos, G., Regina, A., Da Rocha, C., 2005. A computerized infrastructure for supporting experimentation in software engineering. In: ESELAW 2005. URL: https://bit.ly/3yBhh8R.

Miedema, F., 2022. Open Science: The Very Idea. Springer Dordrecht.

Munafò, M., Nosek, B., Bishop, D., Button, K., Chambers, C., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E., Ware, J., Ioannidis, J., 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1 (1), 1–9. http://dx.doi.org/10.1038/s41562-016-0021, 2017 1:1.

NASEM, 2022. National academies for sciences, engineering, and medicine. https://bit.ly/3azwJuv.

Nielsen, M., 2012. Reinventing Discovery: The New Era of Networked Science. Princeton University Press.

Open Science IEEE Computer Society, 2020. The ad hoc committee on open science and reproducibility chapter. https://bit.ly/3P7272e.

OpenAIRE, 2023. Zenodo. URL: https://www.zenodo.org/.

OSF, 2023. OSF repository. https://osf.io/.

Paltenghi, M., Pradel, M., 2021. Thinking like a developer? Comparing the attention of humans with neural models of code. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering. ASE.

Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S., 2007. A design science research methodology for information systems research. J. Manage. Inf. Syst. 24 (3), 45–77.

Peitek, N., Apel, S., Parnin, C., Brechmann, A., Siegmund, J., 2021. Program comprehension and code complexity metrics: An fMRI study. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering. ICSE, pp. 524–536. http://dx.doi.org/10.1109/ICSE43902.2021.00056.

Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M., 2008. Systematic mapping studies in software engineering. In: 12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12. pp. 1–10.

Popper, K., 1959. The Logic of Scientific Discovery. Hutchinson & Co, pp. 88–94.

Prechelt, L., 2011. Plat_Forms: A web development platform comparison by an exploratory experiment searching for emergent platform properties. IEEE Trans. Softw. Eng. 37 (1), 95–108. http://dx.doi.org/10.1109/TSE.2010.22.

Pucci, J., Passos, L., García, R., Messias, R., Olivete, C., 2014. ExpTool : a tool to conduct , package and replicate controlled experiments in software engineering. In: Proceedings of the International Conference on Software Engineering Research and Practice. SERP.

Ralph, P., Baltes, S., Bianculli, D., Dittrich, Y., Felderer, M., Feldt, R., Filieri, A., Furia, C.A., Graziotin, D., He, P., et al., 2020. ACM SIGSOFT empirical standards.

Research.com, 2023. Best computer science journals for software engineering & programming. https://bit.ly/3XfAKYa.

Revoredo, K., Djurica, D., Mendling, J., 2021. A study into the practice of reporting software engineering experiments. Empir. Softw. Eng. 26 (113), http://dx.doi.org/10.1007/s10664-021-10007-3.

Ricca, F., Torchiano, M., Leotta, M., Tiso, A., Guerrini, G., Reggio, G., 2018. On the impact of state-based model-driven development on maintainability: A family of experiments using UniMod. Empir. Softw. Eng. 23 (3), 1743–1790, URL: https://doi.org/10.1007/s10664-017-9563-8.

Robillard, M., Nassif, M., McIntosh, S., 2018. Threats of aggregating software repository data. In: Proceedings - 2018 IEEE International Conference on Software Maintenance and Evolution. ICSME 2018, IEEE, pp. 508–518. http://dx.doi.org/10.1109/ICSME.2018.00009.

Roper, M., Wood, M., Miller, J., 1997. An empirical evaluation of defect detection techniques. Inf. Softw. Technol. 39 (11), 763–775. http://dx.doi.org/10.1016/S0950-5849(97)00028-1.

Runeson, P., Stefik, A., Andrews, A., 2014. Variation factors in the design and analysis of replicated controlled experiments: Three (dis)similar studies on inspections versus unit testing. Empir. Softw. Eng. 19 (6), 1781–1808. http://dx.doi.org/10.1007/S10664-013-9262-Z.

Santos, A., Gómez, O., Juristo, N., 2018. Analyzing families of experiments in SE: A systematic mapping study. IEEE Trans. Softw. Eng. 46 (5), 566–583.

Santos, A., Vegas, S., Oivo, M., Juristo, N., 2021. A procedure and guidelines for analyzing groups of software engineering replications. IEEE Trans. Softw. Eng. 47 (9), 1742–1763. http://dx.doi.org/10.1109/tse.2019.2935720, arXiv:2004.05332.

Sayyad Shirabad, J., Menzies, T., 2005. The PROMISE Repository of Software Engineering Databases. School of Information Technology and Engineering, University of Ottawa, Canada, URL: http://promise.site.uottawa.ca/SERepository.

Shen, B., Zhang, W., Kästner, C., Zhao, H., Wei, Z., Liang, G., Jin, Z., 2021. SmartCommit: A Graph-Based Interactive Assistant for Activity-Oriented Commits. In: ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA, pp. 379–390. http://dx.doi.org/10.1145/3468264.3468551.

Shepperd, M., Ajienka, N., Counsell, S., 2018. The role and value of replication in empirical software engineering results. Inf. Softw. Technol. 99, 120–132.

Shi, L., Chen, X., Yang, Y., Jiang, H., Jiang, Z., Niu, N., Wang, Q., 2021a. A first look at developers' live chat on gitter. In: A First Look At Developers' Live Chat on Gitter. In: ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA, pp. 391–403. http://dx.doi.org/10.1145/3468264.3468562.

Shi, L., Jiang, Z., Yang, Y., Chen, X., Zhang, Y., Mu, F., Jiang, H., Wang, Q., 2021b. ISPY: Automatic issue-solution pair extraction from community live chats. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering. ASE.

Shull, F., Basili, V., Carver, J., Maldonado, J., Travassos, G., Mendonça, M., S., F., 2002. Replicating software engineering experiments: Addressing the tacit knowledge problem. In: Proc. in International Symposium on Empirical Software Engineering. pp. 7–16.

Shull, F., Carver, J., Vegas, S., Juristo, N., 2008. The role of replications in empirical software engineering. Empir. Softw. Eng. 13 (2), 211–218. http://dx.doi.org/10.1007/s10664-008-9060-1.

Siegmund, J., Kästner, C., Liebig, J., Apel, S., Hanenberg, S., 2014. Measuring and modeling programming experience. Empir. Softw. Eng. 19 (5), 1299–1334, URL: https://doi.org/10.1007/s10664-013-9286-4.

Siegmund, J., Siegmund, N., Apel, S., 2015. Views on internal and external validity in empirical software engineering. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering. Vol. 1, IEEE, pp. 9–19.

Solari, M., 2011. Laboratory Package Proposal for Software Engineering Experiments (in Spanish) (Ph.D. thesis). Politécnica de Madrid University, Madrid, pp. 13–16, URL: https://bit.ly/3GGXNlP.

Solari, M., Vegas, S., 2006. Classifying and analysing replication packages for software engineering experimentation. In: 7th International Conference on Product Focused Software Process Improvement (PROFES 2006)-Workshop Series in Empirical Software Engineering (WSESE). Amsterdam, Paises Bajos.

Solari, M., Vegas, S., Juristo, N., 2018. Content and structure of laboratory packages for software engineering experiments. Inf. Softw. Technol. 97 (December 2016), 64–79. http://dx.doi.org/10.1016/j.infsof.2017.12.016.

Springer, 2023. The fundamentals of open access and open research. https://bit.ly/44CkuE4.

SSRN Elsevier, 2022. Open access preprint online community. https://bit.ly/44CB1rE.

Timperley, C.S., Herckis, L., Le Goues, C., Hilton, M., 2021. Understanding and improving artifact sharing in software engineering research. Empir. Softw. Eng. 26 (4), http://dx.doi.org/10.1007/s10664-021-09973-5.

Trisovic, A., Lau, M.K., Pasquier, T., Crosas, M., 2022. A large-scale study on research code quality and execution. Sci. Data 9 (1), 60, URL: https://bit.ly/3DuiJgo.

TSE, IEEE transactions on software engineering journal. https://bit.ly/3Q84cff.

UNESCO, 2021a. Recommendation on open science. https://bit.ly/3rMWhg0.

UNESCO, 2021b. The components of open science. https://bit.ly/44CkP9O.

Uyaguari, F., Guevara-Vega, C., Quiña-Mera, A., Uyaguari, A., Acosta, C., 2023. Effect of requirements specification using native language on external software quality. IET Softw. 17 (3), 287–300. http://dx.doi.org/10.1049/sfw2.12124.

Vegas, S., Juristo, N., Moreno, A., Solari, M., Letelier, P., 2006. Analysis of the influence of communication between researchers on experiment replication. In: ISESE'06 - Proceedings of the 5th ACM-IEEE International Symposium on Empirical Software Engineering. Vol. 2006, pp. 28–37. http://dx.doi.org/10.1145/1159733.1159741.

Wang, Z., Zhang, H., Chen, T., Wang, S., 2021. Would You like a Quick Peek? Providing Logging Support to Monitor Data Processing in Big Data Applications. In: ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA, pp. 516–526. http://dx.doi.org/10.1145/3468264.3468613.

Weinberger, D., 2014. Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room is the Room. Basic Books, Inc., USA.

Winter, S., Timperley, C.S., Hermann, B., Cito, J., Bell, J., Hilton, M., Beyer, D., 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. pp. 145–156.

Woelfle, M., Olliaro, P., Todd, M., 2011. Open science is a research accelerator. Nat. Chem. 3 (10), 745–748. http://dx.doi.org/10.1038/nchem.1149, 2011 3:10.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M., Regnell, B., Wesslén, A., 2012. Experimentation in Software Engineering: an Introduction. Springer.

Wonsil, J., 2021. Reproducibility as a Service (Ph.D. thesis). The University of British Columbia, Vancouver, pp. 13–16, URL: http://dx.doi.org/10.14288/1.0398221.

Zenodo, 2022. Empirical software engineering community on zenodo. https://bit.ly/3PE09a5.

Cathy Guevara-Vega is Associate Professor-Researcher with the Universidad Técnica del Norte. Doctor of Computer Engineering at the University of Seville. She has various publications including scientific articles, book chapters, technical reports in the area of empirical software engineering with focus in human experimentation, requirements engineering, testing and software business process. She is a member of IEEE Computer Society and IEEE WIE Affinity Group. She is a founding member of Network Science Research Group (eCIER). She also serves regularly as a Reviewer for international conferences and high-impact scientific journals.

Beatriz Bernárdez is currently an Assistant Professor with the University of Seville. Her current research interests include empirical software engineering, requirements engineering, and impact of human factor at Software Engineering. She was in collaboration with some international conferences such as ESEM 2007 and SPLC 2017. She has served as a Reviewer for the IEEE Transactions of Software Engineering and Empirical Software Engineering Journal.

Margarita Cruz is currently an Associate Professor with the University of Seville, where she was teaching databases for more than 30 years. Her current research interest includes empirical software engineering, specifically in methodological aspects of experiment replications such as the specification of changes between replications and its reporting not only in software engineering but also in other research areas.

Amador Durán is currently an Associate Professor in software engineering with the University of Seville. His current research interests include requirements engineering, software variability, empirical software engineering, business process modeling, formal methods, and metamorphic testing. He is the author of the REM requirements management tool used by universities and companies in various countries. He also serves regularly as a Reviewer for international journals and conferences.

Antonio Ruiz-Cortés is a Full Professor and elected member of the Academia Europaea. He received the Most Influential Paper of SPLC (2017), VAMOS award (2020), and National Award on Informatics (ARITMEL Award, SCIE, 2022). He leads the Unit of Excellence SCORE and the research group Applied Software Engineering. His research focuses on service-oriented computing, AI for software and service engineering, intelligent contract management with coevolutionary hybrid intelligence, and quantum software engineering. He is an associate editor of IEEE Transactions on Service Computing and Springer Computing and formerly of the International Journal on Cooperative Information Systems. Contact him at aruiz@us.es

Martin Solari is currently an Associate Professor of Software Engineering, Faculty of Engineering, Universidad ORT Uruguay. Ph.D. in Software and Systems (Universidad Politécnica de Madrid). Systems Engineer (Universidad ORT Uruguay). Researcher grade 3 PEDECIBA Informatics. Researcher level I National System of Researchers, Uruguay.