

On Artifacts’ evaluation and the CDUR evaluation protocols

Teresa Gomez-Diaz (University Gustave Eiffel-LIGM-CNRS, France),

Tomas Recio (University Antonio de Nebrija, Spain)

Contact: Teresa.Gomez-Diaz@univ-eiffel.fr, trecio@nebrija.es

27th September 2026

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0), see <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Abstract

Summary. The goal of this article is to adjust the CDUR assessment protocols, initially proposed to evaluate Research Software, and then adapted to Research Data, to an actually quite widespread setting, the one related to scientific outputs labeled as Artifacts.

Background. Understanding Artifacts and their current role in research evaluation is a challenging and relevant issue, still not sufficiently addressed by the scientific community.

Methods. Our previous work on Research Software, Research Data, and the CDUR evaluation protocols, provide the basis to get insights in this new context.

Results. After concluding with a precise definition of Research Artifacts, we will show how the CDUR protocols can be reformulated to apply quite straightforwardly to Research Artifacts evaluation. Moreover, we will also describe how these CDUR protocols could improve the ACM badges and other protocols often used in several research communities for Artifact evaluation.

Conclusion. Critical assessment is a fundamental stage in science, as well as the validation and the reproduction of experimental results. The CDUR protocols provide the means to facilitate the adoption of clear, flexible and transparent Artifact evaluation procedures that enable validation and reproducibility of research results obtained with Research Artifacts, as well as to facilitate their reuse.

1 Introduction

Evaluation plays an important role and shapes a significant portion of the activities in every scientific area. As remarked in [18]:

To produce knowledge as well as is possible, identifying worthwhile interlocutors that can be part of a fruitful conversation is fundamental. Furthermore, the resulting work has to be put in a stable form so as to facilitate its critical assessment. In the print world, “articles” or “monographs” play this role. Communities of trusted members form around them and feed further conversations, further debates. Thus is the “Great Conversation” of science built up.

Nevertheless, evaluation policies and practices can vary deeply according to the key actors, the institutions, the scientific communities, and the different evaluation contexts where these actors are involved in.

This great conversation do produce and fix new knowledge that becomes trusted knowledge, process that is now to be placed in the current digital era. It was still centered in the production of articles, monographs and other publications, and also in conference venues and their related publications, but other digital objects

are becoming more and more relevant in this evolving landscape, like the production of software, data and, more generally speaking, other *artifacts*¹. It is then important to understand their role to establish new knowledge as verified, and adding, thus, new elements to the great conversation.

In science, evaluation is carried out by *peers* and this is a distinctive feature of the scientific activity. Critical assessment is then a fundamental stage of the great conversation building, as well as the validation and the reproduction of experimental results [1]. The question here is how outputs like software, data, and maybe others do participate to the great conversation, since, and as far as we know, there is not a real publication system for these objects that is so largely adopted as the one associated to articles, which is, in our opinion, one of the reasons for their somehow unclear role in order to include them as elements of this great conversation building.

As a first step to study their role in the scientific evaluation ecosystem, we need to understand better these objects as scientific productions. In our previous work we have studied in detail two kind of research outputs: Research Software (RS) and Research Data (RD) from different points of view, as digital objects, as research outputs, but without forgetting some relevant legal aspects [9, 10, 11, 13, 14, 16, 17]. In order to study their role in the great conversation building, we have proposed the CDUR evaluation protocols [10, 14], where CDUR stands for (C) Citation, (D) Dissemination, (U) Use, (R) Research, which are the four steps that compose the protocol and that are to be applied in this chronological order.

The present article is devoted to study how this work on RS, RD and the CDUR protocols extends to a more general framework, the one related to *research artifacts*, to better understand this kind of objects, and to study the applicability of the CDUR protocols for their evaluation.

Artifacts are objects studied in the context of the Experimental Computer Science and Engineering (ECSE) community and, in particular, in other communities related to software engineering [1, 2, 3, 4, 6, 7, 8, 19, 20, 21, 22, 23, 24, 27, 28, 29]. In this work we will adopt the artifact concept that goes on the line of the vision provided in [7] by the ECSE community (see section 2):

An artifact in ECSE is an instance or implementation of one or more computational phenomena. Because the phenomena being studied are processes, algorithms, mechanisms, and the like that manipulate and transform information, the artifact will embody such manipulations and transformations. Examples of artifacts are hardware systems (such as computers) or software systems (such as text editors).[...] It often embodies a substantial portion of the intellectual contribution of experimental research, and its creation represents a significant intellectual effort.

The *Artifact Evaluation (AE) process*² appears first in the joint meeting of the European Software Engineering Conference and the Association for Computing Machinery Special Interest Group on Software Engineering (ACM-SIGSOFT) Symposium on the Foundations of Software Engineering (ESEC/FSE) 2011 conference [20] in order to help evaluate submitted work. It is open to artifacts associated to accepted papers in the conference. Artifacts are classified in two sections, as reported in the ESEC/FSE 2011 AE guidelines: code (i.e. *software*) and non-code artifacts, that includes *data*, to be presented preferably in open document formats.

The present work is then dedicated to better understand artifacts and how they are connected to our previous work on RS and RD. We will also study the AE context and highlight its links with the CDUR evaluation protocols. Its outline is as follows. The next section is devoted to understanding the concepts of artifacts used in the consulted bibliography. Then, in section 3, we will focus on two of the most relevant artifacts, RS and RD that we have studied in our previous work, work that we will use as the basis to provide a definition of Research Artifacts in section 3. Section 4 is dedicated to AE, ACM badges and other initiatives, and it will be followed by a section on the CDUR evaluation protocols and its application to this more general context of artifacts. Conclusions end this work.

Please note that this work is placed in a purely academic context, but the propositions of CDUR can be studied and adapted to industrial and non academic contexts.

¹Please note that both spellings, *artifact* and *artefact* are used in the consulted bibliography. In what follows we use the word *artifact* except when referring to the literature that uses the word *artefact*. The Annex I includes some dictionary consultations on this concept.

²<http://2011.esec-fse.org/cfp-artifact-evaluation>

2 Understanding Artifacts

The work [20] refers to this definition from [23]:

Definition 2. *An artefact is a self-contained work result, having a context-specific purpose and constituting a physical representation, a syntactic structure and a semantic content, forming three levels of perception.*

i.e. *describing an artefact as a self-contained work result with a physical representation and a context-specific purpose.* The analysis carried out in [20] mentions diverse *types of research artifacts*: code and software, data, proofs, but also grammars, surveys, and even hardware, and it studies more in detail three types of *research artifacts*: code, proofs (for example for the proof assistants Coq, Isabelle...) and data.

We have found these considerations in [23] in order to propose the artifact definition:

Software development projects, especially when being complex, embody a multitude of artefacts that are developed, used, maintained, evolved, and which depend on each other. The final software product as such constitutes an artefact as well as the models of the software product, its specifications, safety and security certificates, project management plans, user manuals, the code, and many more. [...] the question arises if all these are to be seen as one artefact or if they constitute different artefacts, and if so, how we can describe their relationship.

and that, regarding the physical representation,

artefacts can appear in different representations, for instance as files on a computer, as elements of data bases, or as printed documents [...]

or that

we can observe that different research areas came up with their own definition of what constitutes an artefact in order to examine or transform it at certain levels of perception.

considerations that drive to the above cited definition [20, 23]. Moreover, in [20] we can find:

We found different expectations depending on the artifact type. Although documentation is mentioned for all three artifact types, especially for proof and code artifacts, there are different expectations, such that code requires documentation, and proofs require completeness and understandability. This is not surprising, because program code can be supplied in multiple forms and languages, whereas mechanized proofs are usually formulated using one of the major proof assistants (i.e., Coq, Isabelle, etc.).

Other works refer to artifacts as *code and non-code artifacts* (Guidelines for ESEC/FSE 2011 AE process submissions), as *tools, benchmarks, data, and more* [27] or as *prototypes*, i.e. *code, and data (the research artifacts)* [29]. Also in [29], authors refer to the ACM Task Force on Data, Software, and Reproducibility in Publication³ and they provide this definition from [1] (see also [19]):

By “artifact” we mean a digital object that was either created by the authors to be used as part of the study or generated by the experiment itself. For example, artifacts can be software systems, scripts used to run experiments, input datasets, raw data collected in the experiment, or scripts used to analyze results.

Moreover, the ACM provides a *Software Artifacts Collection*⁴ and a *Data Artifacts Collection*⁵ among other Collections⁶, and the preceding artifact definition is proposed for both *software* and *dataset* objects in these two collections. We have also consulted The Dagstuhl Research Online Publication Server (DROPS)⁷

³<https://www.acm.org/publications/task-force-on-data-software-and-reproducibility>

⁴<https://dl.acm.org/artifacts/software>, with 2873 artifacts in September 2026.

⁵<https://dl.acm.org/artifacts/dataset>, with 6 artifacts available in September 2026.

⁶<https://dl.acm.org/collections>

⁷<https://drops.dagstuhl.de/>

which provides a collection of *Supplementary Materials*⁸. We can conclude that these two databases (ACM, DROPS) include mainly software, few databases and rarely other kind of artifacts. As pointed out previously, other works also confirm that the considered artifacts are, in most of the cases, software and/or data produced by the authors of the related (and accepted) conference papers. It seems to us that the use of the word artifact has the goal of *not to forget some kind of output that can be associated to the accepted papers*, but that, in most of the cases, it refers mainly to software, and secondly to data.

Several times we have found in the consulted bibliography that the artifact descriptions do make differences between code, software, scripts, or proofs, the later being, as far as we understand, scripts for proof assistants. Note that we do not discriminate here between scripts, code or software, and that we do not consider documentation as an independent artifact, but part of the software or of the data, as explained in the next section. This last point is confirmed by the guidelines we have consulted [1, 2], and in those of the ESEC/FSE 2011, as usually the first requirement is that the artifact provides good and clear documentation, and this documentation is not usually considered an artifact in its own [19, 20, 22]. On a similar line, we will not distinguish between data, datasets, databases, input or output data.

Furthermore, and as remarked in [23], the perception on artifacts seems to vary upon the different communities. As announced in the introduction, in this work we will adopt an artifact concept that goes on the line of the vision provided by the ECSE community [7]:

An artifact in ECSE is an instance or implementation of one or more computational phenomena. Because the phenomena being studied are processes, algorithms, mechanisms, and the like that manipulate and transform information, the artifact will embody such manipulations and transformations. Examples of artifacts are hardware systems (such as computers) or software systems (such as text editors).[...] It often embodies a substantial portion of the intellectual contribution of experimental research, and its creation represents a significant intellectual effort.

where it is remarked as well that

In general, when the computational phenomenon is complex, an artifact that embodies the idea will also be complex and will have many component parts.

A much more detailed and complete view of the concept of artifact can be found in [7] (p.16), see also its section “More on Artifacts” (p.28), and further considerations in the same line of thought can be found in “The IT Artifact” section of [6].

As Artificial Intelligence (AI) becomes an integral part of almost every Information Technology (IT) we use both in our private lives and at work today, it is necessary to distinguish AI from the IT artifact [6]. Thus, another kind of artifacts, *AI artifacts* is becoming part of this landscape.

In [6], AI artifacts are studied from three groups of theories on intelligence. For example, from the *Perspective of General Intelligence*, AI artifact refers to

intelligent machines capable of performing tasks and sensing their environment. These capabilities should be measurable and testable. That is, we should be able to distinguish between two AI artifacts based on their capacity to do work (e.g. their computational abilities and speed at which they deliver expected results) and their ability to detect changes in their environment (e.g. voice command, temperature change...).

and further considerations are provided following the *Perspective of Cognitive Intelligence* or the *Perspective of Cognitive Intelligence*. Among the differences studied between AI and IT artifacts in [6], we highlight the following ones.

- *The same AI artifact you use today would behave differently at a different point in time under the same or different circumstances or contexts without being reprogrammed by man.*
- *AI artifacts can adapt to their users.*

As the main subject of this work, AE placed in the context of facilitate artifact reuse and reproducibility in experimental science (see section 4), AI artifacts fall somehow outside the scope of this work. Maybe further research is needed in order to understand better this AI artifact context, see for example the research agenda of Prompts as Software Engineering Artifacts [28].

⁸<https://drops.dagstuhl.de/entities/collection/supplementary-materials>, this collection includes 378 *Software*, 15 *Dataset* and some other materials in September 2026.

3 Research Software, Research Data and Research Artifacts

In [10, 13] we have provided the following definitions for Research Software (RS) and Research Data (RD), definitions that are also included here in order to facilitate the reading of this work.

Research Software *is a well identified set of code that has been written by a well identified research team. It is software that has been built and used to produce a result published or disseminated in some article or scientific contribution. Each RS encloses a set of files containing the source code and the compiled code. It can also include other elements as the documentation, specifications, use cases...*

Research Data *is a well identified set of data that has been produced (collected, processed, analyzed, shared & disseminated) by a research team. The data has been collected, processed and analyzed to produce a result published or disseminated in some article or scientific contribution. Each RD encloses a set (of files) that contains the dataset maybe organized as a database, and it can also include other elements as the documentation, specifications, use cases, and any other useful material as provenance information, instrument information... It can include the RS that has been developed to manipulate the dataset (from short scripts to RD of larger size) or give the references to the software that is necessary to manipulate the data (RS or other).*

These objects, these research outputs are to be understood through three different lens: as *digital objects*, as *research productions*, and as items that are produced by some team and that are maybe to be shared and disseminated, and thus *contracting a legal status*.

The above definitions fix their status as digital objects i.e. with a physical representation, as already remarked in [23] (see section 2), as they are constituted by a set of files. Note that, here, we identify a dataset organized as a database as a (maybe big) file or set of files, that is, the basic structure manipulated in computers. In our vision, each file is not to be considered as an independent artifact, as it may lose valuable context information without considering it as part of a whole object. In the same line, we also understand here *graphs* as datasets.

The Harvard resource File Naming Conventions⁹ provides valuable information in naming the files and structuring them. Another important aspect is the format in which these files are provided in, and tools like FACILE¹⁰ of the Centre Informatique National de l'Enseignement Supérieur (CINES) in France can be of help to deal with this question. Formats and standards may be different in some scientific communities, and best sharing and dissemination practices take these aspects into account.

The above definitions also fix their status as research productions, as they have been realized in order to answer a research question, to produce a scientific result, which corresponds to the *context-specific purpose* of Definition 2 in [23]. Their research production team should be correctly identified, with names and roles transparently described, as they are the persons producing knowledge and participating in the great scientific conversation. This research team is the one deciding what is to be shared and disseminated (which version, which files...), how, when, where, with whom, under what conditions... see [13, 14, 16, 17].

Note that science is evolving in the direction of considering each research output as a valuable object, and, then, RS and RD have their own identity in the research landscape, and the teams that produce RS, RD or the related research articles maybe slightly different, and surely including different roles and responsibilities [10, 15]. Also note that [4, 7]:

Artifacts are another vehicle for the dissemination of knowledge. [...] The idea that is embodied by the artifact disseminates along with the artifact. Many regard the dissemination of an artifact as an important form of knowledge dissemination that is equivalent to dissemination by publication.

which provides further arguments to value RS and RD as different research productions that can be evaluated in an independent way of the one applied to their related publications.

Whether we like it or not, the dissemination stage of a research output endorses it with a legal status. For a research article (or any written document), it fixes its authors and, thus, its copyright holders. This is

⁹<https://datamanagement.hms.harvard.edu/plan-design/file-naming-conventions>

¹⁰<https://facile.cines.fr>

the same for software, as its main component, the source code, is a written work, but some differences with respect to articles do apply (and have been studied in the French legal context in [9, 11, 17]). In the case of articles, it is usually the time in which authors sign agreements for copyright transfer to publishers. Note that this legal step differs, for example, regarding Anglosaxon or French copyright laws [5].

In our work, the provided RS definition relies over the *software* concept as a legal object [14], with its title, authors and rightholders, which allows us, among other considerations, to avoid giving a lot of importance to the differences between *scripts* and software, as, from the legal point of view, they are protected in the same way. We also consider here that, from this point of view, proofs for proof assistants can be considered in the same line as scripts and, thus, as software. On the other hand, we have found regularly *code* as a synonym of *software*. In fact, we prefer to use the concept of software, and stress the fact that this object includes source code and maybe compiled code, as well as other elements.

On the other hand, *data* is much more difficult to grasp [13], but this is not an obstacle to build a RD definition with a formulation pretty similar to the one given for RS. The legal context of a dataset can be easy and simple to be understood and manipulated, but release a RD can trigger rather complex legal contexts, aspects that have been studied in [13, 14] in order to give the above RD definition. For example, a dataset can be organized as a database, that triggers the *sui generis* database rights in the French context (for example). The Ressources juridiques en droit des données de la recherche¹¹ of the French University of Montpellier can be a tool to help to determine the legal contexts that can be necessary to be taken into account *before* releasing the RD.

Note then that, in our work, data includes datasets, databases, input or output data, generated by software or by other instruments or collected by a research team.

Licenses have the important role of reinforce the legal status of the object [9, 11]. They are the legal documents that should go along with the research outputs, as they establish the conditions under which these objects are shared and disseminated, and indicate whether they may be copied, modified, and/or redistributed. Most popular and well known licenses can be found in the following links: Creative Commons (CC)¹², Free Software Foundation (FSF)¹³, Open Source Initiative (OSI)¹⁴, Software Package Data Exchange (SPDX)¹⁵.

Licenses should be put into place *before* releasing the outputs in order to inform the potential users about their sharing and dissemination conditions, and the decisions regarding these legal aspects should be taken into account as soon as possible in the lifetime of these objects, in order to avoid what can be unexpected an maybe annoying consequences [9, 11].

Regarding the documentation that goes with the research outputs, it could be considered as a whole artifact, independent of its associated RS or RD, as it can have its own set of files, its own author's team, its own license. But it usually has, in our opinion, little value considered as an independent, isolated object, and without being associated to the documented output.

To consider best dissemination practices and Software or Data Management Plans can be of help to deal with important questions that should not being forgotten at the release step [11, 14], works that also include a flexible dissemination procedure for RS and for RD.

Inspired by the RS and RD definitions, and following the ECSE artifact definition, we propose here a definition of Research Artifacts (RA) as follows:

Research Artifact *is an artifact that has been produced by a well identified research team. It is an artifact that has been built and used to produce a result published or disseminated in some article or scientific contribution.*

We classify here RA as digital and no-digital objects. The digital objects are mainly (digital) documents, software, data (including databases), and maybe other kind of objects, and consist of a set of files, usually including a documentation as the one that goes usually with software or data. Therefore RA include RS and RD.

¹¹<https://jurisdor.fr/>

¹²<https://creativecommons.org>

¹³<https://www.gnu.org/licenses/license-list.html>

¹⁴<https://opensource.org/license>

¹⁵<https://spdx.org/licenses/>

Research articles and scientific publications in general, including conference presentations, posters, etc. could be considered in this large and general context, but we prefer to avoid it, as they have their own identity in the scientific landscape, well established since long.

Non digital objects include hardware, and are also usually accompanied with documentation, and maybe software and/or data as well. Computers, robots can be RA if developed and built in a scientific framework. The CERN Open Hardware Repository¹⁶ provides a large set of RA examples, and also mentions the licenses¹⁷ used to release these objects.

Whether other research outputs like chemical molecules, nanomaterials, AI artifacts, etc. are to be included in this RA definition is left to the diverse scientific communities that will weight the specificities and the benefits of understanding these research outputs in the context studied here.

This RA definition provides the context where RA are understood as digital or no digital artifacts, or as multifold objects containing digital and non-digital components. They are also correctly understood as scientific productions of research teams that are associated to research results and scientific contributions. They have a legal context and associated licenses.

4 Artifact Evaluation: ACM badges and other initiatives

As mentioned in [7] (p.35) the evaluation of scholarly work is intimately connected with the artifact in ECSE. This is the kind of evaluation process that has been first implemented in the AE initiative of the ESEC/FSE 2011 conference [20]. It is open to artifacts associated to accepted papers in the conference, where artifacts are classified as code and non-code artifacts. The ESEC/FSE 2018¹⁸ mentions an Artifact Track, where *research papers with “artifacts” receive a (ACM) “badge” on the front page of their paper in the proceedings*. These ACM badges have the following titles: *Functional*, *Reusable*, *Available*, *Replicated*, *Reproduced*, and certify that the paper has associated artifacts that have been evaluated by an independent committee. An example of a paper of the ESEC/FSE 2020¹⁹ conference with three ACM badges is [20].

The ACM badges [1] seem to be the more mentioned in the consulted bibliography and conferences. See the images of the ACM badges, for example, in the ESEC/FSE 2018 and 2020 websites and in [1, 2, 3, 25]. They are resumed here as follows (in their version of August 2020, see also [25]):

- *Artifacts Available*: refers to the artifact being publicly accessible in a repository and having a DOI.
- *Artifacts Evaluated-Functional*: refers to the artifact documentation and includes evidence of verification and validation.
- *Artifacts Evaluated-Reusable*: extends the badge Evaluated-Functional, it refers to the artifact quality (exceeding “functional”) and includes considerations on the documentation quality, facilitates reuse, and adheres the community norms and standards.
- *Results Replicated*: external teams can obtain the same results using, in part, artifacts provided by the paper authors.
- *Results Reproduced*: external teams can achieve independently the same results as the ones presented in the paper, without the use of author-supplied artifacts.

It is now an established review practice in several fields of computer science research [1, 2, 19, 25, 29], see for example the websites of the ACM International Conference on the Foundations of Software Engineering (FSE)²⁰, the Proceedings on Privacy Enhancing Technologies symposia²¹ or the Systems Research Artifacts website²² that collects artifact evaluation results for several computer systems conferences (ATC, CAIS, EuroSys, FAST, OSDI, SC, SOSP). The ACM badges used in the 43rd International Conference on Software

¹⁶<https://ohwr.org/>

¹⁷<https://ohwr.org/licences/>

¹⁸<https://2018.fseconference.org/track/fse-2018-Artifacts>

¹⁹<https://2020.esec-fse.org/track/esecfse-2020-artifacts>

²⁰<http://www.esec-fse.org/>

²¹<https://petsymposium.org/popets/>

²²<https://sysartifacts.github.io/>

Engineering (ICSE'21) are explained in [2]. A decade of artifact evaluation and the use of these badges has been analyzed in software engineering and programming languages conferences [29], and an extended model has been introduced in Brazilian conferences on networking and distributed systems [19]. The benefits of AE introduction are studied in the security field [21], and it is also the subject under study in the recent Dagstuhl Seminar “Towards Standards for Research Artifact Sharing and Open Science in Software Engineering Research”²³. In [22] it is investigated ten years of ICSE conference papers with associated software artifacts as well as a large context of artifact publishing guidelines that include the ACM badges. Other badgering initiatives proposed by several organizations, scientific publishers and scientific journals can be found, for example, in [25].

This AE initiative introduces a somehow *new* evaluation context where authors of the articles accepted to the conference are invited to submit the associated artifact in order to promote the *integrity of the research ecosystem* [1], *to actively engage with the community in order to support them in making their research artifacts publicly available and in fostering replication of research results* [2], to improve the *reproducibility of research* presented in the accepted papers [29], to enhance *reuse* [20], and to *increase the quality of the submitted papers and artifacts*, as well as to contribute to the *continuous improvement of the scientific community* as a whole [19]. All this because *an experimental result is not fully established unless it can be independently reproduced* [1]. All in all, this shows that a variety of research communities and different kind of organizations have embraced AE with the goals of reproducibility in experimental science and to enhance the reuse of artifacts, among others.

The terms *Repeatability* (same team, same experimental setup), *Reproducibility* (different team, same experimental setup), *Replicability* (different team, different experimental setup) are defined by the ACM [1] and appear regularly in the consulted artifact related bibliography, as well as *validation* and *reuse*, see, for example [1, 2, 19, 20, 21, 24, 25]... where [19] also mentions *falsifiability*.

It is our claim that the proposed CDUR protocols [10, 14] provide mechanisms to complete and to improve the ACM badges for RA evaluation, as explained in the next section.

5 Evaluating Research Artifacts: the CDUR protocols

We have proposed the CDUR evaluation protocols first for RS [10], which has been achievable because of our experience and deep understanding of this research output [9, 10, 11, 15, 16, 17]. Then we have studied the context of RD [13] and show that the CDUR protocols are valid as well in this context [14]. For this we have explained how the application of the same framework protocol for RD is adequate, and that one of the forces of CDUR is then that it is easy to be adapted to other research outputs than RS. The fact that CDUR protocols have been mentioned in the work to tailor the TRAAC tool proposed to evaluate a large variety of state-of-the-art tools and methods in the context of nanomaterials [26] contribute to show the strengths and the flexibility of these protocols.

We include here a short introduction of the CDUR protocols for RS or RD [10, 14]. It is composed of four steps to be applied in this chronological order:

(C) Citation. This step measures if the RS or RD are well identified as a research output, i.e. if there is a good citation form, or good metadata. We look also here for best citation practices.

This is a legal related point where we ask for authors (if any) and for the members of the research production team to be well identified, which are their affiliations, and, for example, the % of their participation in software writing or which is their role in the output production.

(D) Dissemination. In this point we look for best dissemination practices, in agreement with the scientific policy of the evaluation context. The dissemination of RS and RD needs a license to set the sharing conditions. This is also the step dealing with legal (and *ethical*) issues, for example there are maybe further legal issues in RD that should be considered before releasing this object (personal data, *sui generis* database rights...).

This is a policy point in which we look at Open Science requirements [12, 24].

²³<https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/26381>

- (U) **Use.** This point examines “software” or “data” aspects, in particular the correct results that have been obtained with the RS or RD, and we can also look if their reuse has been facilitated, the output quality, best software/data practices such as documentation, testing, installation or reuse protocols, up to read the code, launch the RS, use examples...

This is the reproducibility point that looks at the validation of scientific the results obtained with the RS and/or the RD.

- (R) **Research.** This point examines the research aspects associated to the RS and/or RD production: the quality of the scientific work, the proposed and coded algorithms and data structures, which are the related publications, the collaborations involving these outputs, the funded projects...

This point measures the impact of the RD and/or RS related research.

Note that the CDUR protocols are proposed as a help to evaluation committees and decision makers [10], as they are the key evaluation actors that will adapt these steps to a particular evaluation case, as well as to decide the different step’s weight and the balance among them that is to be applied, since evaluation policies and practices can vary deeply in different evaluation contexts and scientific communities. CDUR protocols can be also of help to evaluated researches in order to know how to proceed facing an evaluation setting, in the case the used evaluating protocol is available beforehand, which can place the evaluation context in a very transparent framework.

As studied in [14], there is no modification of the basic structure of the CDUR protocols when applied to the RD case, but some adaptations are necessary. The *Citation step* should consider the citation form of the RD, best citation practices for other RD (or RS...), and consider the legal issues related to authorship and copyright (if any). But, as explained in section 3, the legal context for RD can be complex, which should be studied by the producer team before their release, point that we have included to be evaluated in the *Dissemination step*. The initial CDUR protocol concentrates all the aspects related to software in the *Use step*, which was very fortunate, as this step remains the same for RD, and, so, easy to adapt to data related quality issues. The *Research step* remains the same for RS and for RD.

Therefore, to apply the CDUR protocols for RA is pretty much similar as to apply them to RD: the legal issues specific for the artifact are to be studied in the *Dissemination step* and the *Use step* will consider, as well, the quality issues related to the specific artifact in hands. The *Citation step* is where authorship and copyright issues are to be considered, as they are important to set the citation form, and the *Research step* remains also the same as for RS and RD. These steps are to be adapted to the different practices of diverse scientific communities.

The CDUR four steps are related with the ACM badges [1] as follows. The *Artifacts Available* refers to points to be considered in the CDUR *Dissemination step*. The other ACM badges *Artifacts Evaluated-Functional*, *Artifacts Evaluated-Reusable*, *Results Replicated*, *Results Reproduced* are all related to the CDUR *Use step*, as they refer to different levels of quality issues related to the RS or RD, and it is also the step associated to validation and reproducibility issues.

The CDUR steps are also related to the “Five recommendations for FAIR software”²⁴. For example, its first recommendation *Use a publicly accessible repository with version control*, is related to the ACM badge *Artifacts Available* and is one of the points to be considered in the CDUR dissemination step, which also includes the use of licenses for appropriate dissemination and the use of a community registry. The other recommendations refer to the CDUR citation and use steps.

On the other hand, the CDUR protocols improve the ACM badges as follows. The badges do not discriminate if they are considering quality aspects related to software, data, or other kinds of artifacts [20], while the CDUR *Use step* is to be adapted by the committees to the output they are evaluating. Moreover, neither the *Citation* nor the *Research* steps seem to appear in the badges, as these points are associated to the conference papers: it is the conference paper that will be cited, and it is the conference accepting procedure that will evaluate the research issues. But this also means, for example, that the citation mechanisms for RA are neither evaluated nor improved.

The four badges studied in [19] are a variation of the ACM badges. The first one, *Artifact Available* refers much more in detail to dissemination issues, and asks for an artifact (code and/or data) available in some kind of repository with a `README.md` file, and approaches ten different items, including the RA title, a

²⁴<https://fair-software.eu/>

goal description, a description of the running environment, etc. These ten points include also items referring to facilitating artifact reuse that we have included in the CDUR *Use step*, where we can include as well the three other badges: *Artifact Functional*, *Artifact Sustainable*, *Artifact Reproducible*.

Citations and research impact are discussed in [8] (and the mentioned references there) in the context of the Conference ICSE 2009, the 31st International Conference on Software Engineering. We can find there the following considerations:

- *different stakeholders need to evaluate research* (p.35)
- *numbers are simplistic and misleading* (p.35)
- *Sole reliance on citation data provides incomplete and shallow understanding of research, only valid if reinforced by other judgments* (p.36)
- *Citation data and statistics provide some valuable information, [...] but only a limited and incomplete view of research quality* (p.43)
- *Expert judgment cannot be eliminated* (p.43)
- *Research too important to measure only with a coarse tool* (p.43)
- *Favor research products backed by tools, but verify that the claimed results of their use can be replicated* (p.54)
- *Reward successful reuse of one's research rather than paper citation* (p.54)
- *we can do more and better... Towards impact-aware research?* (p. 56)

On the other hand, in [7] (p.68) we can find that:

Because an artifact often embodies aspects of the intellectual content of ECSE research that may be intangible, it is important to consider how this content is communicated to the research community.

and that

How can impact be measured? The principle underlying impact in ECSE is simple: an artifact or an idea has impact if it changes the way other people work.

It is our claim that the CDUR proposed protocols provide a complete framework where the above mentioned issues can be taken into consideration by the evaluation committees, as they rely on expert judgment and not only in measures that can be coarse. CDUR can of be help to understand, measure and improve the impact related knowledge, usually specific to a scientific community, providing thus better assessment methods. Committees involved in the application of CDUR protocols can be composed by different stakeholders representing different levels of expertise regarding different scientific areas, but also different competences in specific RA technical issues.

Note, for example, that the *Citation step* is the first one proposed in CDUR protocols. If the goal of the scientific communities is that RS, RD, RA and any other kind of research outputs acquire value as independent outputs, it is necessary to adopt referencing and citation mechanisms and practices, such an important issue in the case of publications, as the *visibility* of an output is usually analyzed in the terms of *citation* issues, which needs standardized ways to reference them [29]. On the other hand, citations are also a way to measure scientific impact [4, 8, 22].

The CDUR protocols help to put these mechanisms in context with respect to research impact, Open Science policies and other quality issues. On the other hand, the *Use step* of CDUR connects *reproducibility*, *replicability* and *validation* with output's *reuse*, and it is up to the evaluation committee to set the different levels that are to be positively evaluated.

Another point where CDUR protocols can be useful is to facilitate the role of reviewers that will participate in the artifact evaluation, as they can help the conference organizers to decide levels, step's weights and

balance, standards... before hand, and to better clarify the evaluation frame that is to be put into place and applied by the reviewers, as it may happen that the lack of such quality standards leave reviewers without guidance [20, 21], see also the difficulties commented in Moritz Beller’s blog²⁵.

6 Conclusions

The ESEC/FSE 2011 conference has introduced an evaluation context where articles accepted to the conference can get badges related to the evaluation of associated artifacts. This kind of evaluation has been extended to many other conferences and research contexts, with the main goals of improving and facilitating the reproducibility of the results presented in the related articles and the reuse of the artifacts.

In this paper we have analyzed the concepts of artifacts as well as this AE context, evaluation that we have placed as part of the *Great Conversation* that builds new, trusted knowledge in science.

We have recollected previous work on RS and RD, and applied it as the basis for the proposed RA definition. Then we have shown how the CDUR evaluation protocols built initially for RS, and then applied to RD, are easily applied as well to RA evaluation. We have also shown the relationships between CDUR and the ACM badges for AE.

The AE badges are applied to the conference papers, not to the artifacts, and thus, this process dissociates somehow the evaluation of the artifacts and the one of the corresponding papers. But the quality of artifacts is as essential as that of the article itself [19] and researchers do more than write papers [3]. This means in a way that the artifacts are not recognized as valuable as the associated conference articles, as they are evaluated in a secondary role, and only if their authors are willing to pass another evaluation step. This also means that these AE frameworks can be improved in order to recognize fully the artifact’s fundamental role in the complete understanding of a scientific article, as they add a practical dimension that enriches the academic content [19]. It is also necessary to acknowledge that the dissemination of an artifact can be an important form of knowledge dissemination, equivalent as the dissemination by publication [4, 7]. Moreover, as the creation and maintenance of artifacts are not well integrated in the AE system, artifacts are likely to escape the common citation tracking mechanisms [29].

All these considerations confirm the need of evaluation processes adapted to the artifacts, flexible enough to be adjusted easily to different kind of artifacts, as well as to the usual practices of different research communities and evaluation settings. These mechanisms should enhance the value of these outputs, foster their reuse, and recognize their part in the research work essential to the knowledge production usually disseminated by publications: without the RA, the creation of this new knowledge would be much more difficult or perhaps impossible.

We have shown that the CDUR protocols deal correctly with these issues. They provide the means to facilitate the adoption of clear, flexible and transparent evaluation procedures. They enable validation and reproducibility of research results published in the RA related articles, and facilitate RA reuse by enhancing the evaluation standards that will be considered by the community experts for their evaluation, standards that will take into account citation, dissemination, quality and research impact.

This is an important step for the recognition of RA production and, therefore, to help scientists towards better, more efficient research.

²⁵<https://inventitech.com/blog/why-i-will-never-review-artifacts-again/>

Annex I. Dictionaries

In this section we show the consulted dictionaries in order to understand better the concept of *Artifact* from different points of view. Complementary information can be found in the online encyclopedia Wikipedia²⁶

[EN] Collins Dictionary ²⁷

artifact in American English
noun

1. any object made by human beings, esp. with a view to subsequent use
2. a handmade object, as a tool, or the remains of one, as a shard of pottery, characteristic of an earlier time or cultural stage, esp. such an object found at an archaeological excavation
3. any mass-produced, usually inexpensive object reflecting contemporary society or popular culture artifacts of the pop rock generation
4. a substance or structure not naturally present in the matter being observed but formed by artificial means, as during preparation of a microscope slide
5. a spurious observation or result arising from preparatory or investigative procedures
6. any feature that is not naturally present but is a product of an extrinsic agent, method, or the like statistical artifacts that make the inflation rate seem greater than it is

Also: artefact (UK)

[ES] Diccionario de la Real Academia Española (RAE) ²⁸

1. m. Objeto, especialmente una máquina o un aparato, construido con una cierta técnica para un determinado fin. Un artefacto electrónico. Un artefacto volador. Sin.: aparato, máquina, artificio, ingenio, dispositivo, mecanismo.
2. m. despect. Máquina, mueble o, en general, cualquier objeto de cierto tamaño. Sin.: artilugio, trasto, armatoste, mamotreto, cachivache, cacharro.
3. m. Carga explosiva; p. ej., una mina, un petardo, una granada, etc.
4. m. En un estudio o en un experimento, factor que perturba la correcta interpretación del resultado.

[FR] Dictionnaire Larousse ²⁹

artefact

nom masculin

(mot anglais, du latin *artis facta*, effets de l'art.)

(Ce mot d'origine latine emprunté à l'anglais ne prend pas d'accent sur le e : un artefact.)

1. Structure ou phénomène d'origine artificielle ou accidentelle qui altère une expérience ou un examen portant sur un phénomène naturel.
2. Altération du résultat d'un examen due au procédé technique utilisé.
3. En anthropologie, produit ayant subi une transformation, même minime, par l'homme, et qui se distingue ainsi d'un autre provoqué par un phénomène naturel.

²⁶See <https://fr.wikipedia.org/wiki/Art%C3%A9fact> for the French page and <https://es.wikipedia.org/wiki/Artefacto> for the page in Spanish. Software development artifacts have a page in English [https://en.wikipedia.org/wiki/Artifact_\(software_development\)](https://en.wikipedia.org/wiki/Artifact_(software_development)) and in Spanish [https://es.wikipedia.org/wiki/Artefacto_\(dise%C3%B1o_de_software\)](https://es.wikipedia.org/wiki/Artefacto_(dise%C3%B1o_de_software)). These Wikipedia pages are also available in other languages.

²⁷<https://www.collinsdictionary.com/dictionary/english/artifact>

²⁸<https://dle.rae.es/artefacto>

²⁹<https://www.larousse.fr/dictionnaires/francais/artefact/>

ARTEFACT - subst. masc.

A.- Dans les sc. exp. Produit dû à la méthode employée.

B.- P. ext. ‘‘Produit de l’art, de l’industrie’’, objet artificiel

ÉTYMOL. ET HIST. - 1920-24, supra ex. 1. Prob. empr. à l’angl. artefact
‘‘ce qui est réalisé par l’homme, produit artificiel’’ attesté dep. 1821
(Coleridge, Black Mag. X. 256 ds NED t. 13);
formé sur le lat. ars, artis (art*), et le part. passé de facere (faire*).

³⁰<https://www.le-tresor-de-la-langue.fr/definition/artefact>

References

- [1] Association for Computing Machinery (ACM) (2020). Artifact Review and Badging Version 1.1, <https://www.acm.org/publications/policies/artifact-review-and-badging-current>.
- [2] Silvia Abrahão, Daniel Mendez (2021). ICSE'21 Artifact Evaluation Track-Submission and Reviewing Guidelines, <https://conf.researchr.org/getImage/icse-2021/orig/Submission.and.Reviewing.GuidelinesNUEVO.pdf>.
- [3] Maria Teresa Baldassarre, Neil Ernst, Ben Hermann, Tim Menzies, and Rahul Yedida (2023). (Re)Use of Research Results (Is Rampant). *Communications of the ACM* 66, 2, pp. 75-81, <https://doi.org/10.1145/3554976>.
- [4] Richard Baskerville, Abayomi Baiyere, Shirley Gregor, Alan Hevner, and Matti Rossi (2018). Design Science Research Contributions: Finding a Balance between Artifact and Theory, *Journal of the Association for Information Systems*, 19(5), <https://aisel.aisnet.org/jais/vol19/iss5/3>.
- [5] Daniel Battesti and Anne-Laure Stérin (2026). Modèles français et anglo-saxon en matière de licences et de cession de droits : les conséquences pour l'intégration des revues françaises au DOAJ, *Arabesques [Online]*, 120, <https://publications-prairial.fr/arabesques/index.php?id=4548>.
- [6] Ransome Bawack, Samuel Fosso Wamba, Kevin D.A. Carillo (2019). From IT to AI Artifact: Implications for IS Research on AI Adoption and Use, *DIGIT 2019 Proceedings*, 6, <https://aisel.aisnet.org/digit2019/6>.
- [7] Committee on Academic Careers for Experimental Computer Scientists (1994). *Academic careers for experimental computer scientists and engineers*, Washington, DC, National Academy, <https://www.nationalacademies.org/publications/2236>.
- [8] Carlo Ghezzi (2009). Reflections on 40+ years of software engineering research and beyond: an insider's view. In keynote address in the 31st International Conference on Software Engineering, <https://www.cs.uoregon.edu/events/icse2009/keynoteSpeakers/ICSEkeynote.pdf>.
- [9] Teresa Gomez-Diaz (2015). Article vs. Logiciel: questions juridiques et de politique scientifique dans la production de logiciels, 1024 : *Bulletin de la Société Informatique de France*, <https://doi.org/10.48556/SIF.1024.5.119>.
- [10] Teresa Gomez-Diaz, Tomas Recio (2019). On the evaluation of research software: the CDUR procedure [version 2; peer review: 2 approved], *F1000Research* 2019, 8:1353, <https://doi.org/10.12688/f1000research.19994.2>.
- [11] Teresa Gomez-Diaz (2019-26). Les logiciels de la recherche et leurs licences : trois visions sur un objet, *Training support*, V4, University Gustave Eiffel, <https://hal.science/hal-02434287v4>.
- [12] Teresa Gomez-Diaz, Tomas Recio (2020-21). Towards an Open Science definition as a political and legal framework: on the sharing and dissemination of research outputs. Version 2 published in *POLIS N. 19* (2020), <https://doi.org/10.58944/yuro5734>. Version 3 (2021) available on Zenodo <https://doi.org/10.5281/zenodo.4577066>.
- [13] Teresa Gomez-Diaz, Tomas Recio (2022). Research Software vs. Research Data I: Towards a Research Data definition in the Open Science context [version 2; peer review: 3 approved], *F1000Research* 2022, 11:118, <https://doi.org/10.12688/f1000research.78195.2>.
- [14] Teresa Gomez-Diaz, Tomas Recio (2022). Research Software vs. Research Data II: Protocols for Research Data dissemination and evaluation in the Open Science context [version 2; peer review: 2 approved], *F1000Research* 2022, 11:117, <https://doi.org/10.12688/f1000research.78459.2>.
- [15] Teresa Gomez-Diaz, Tomas Recio (2023). Articles, software, data: An Open Science ethological study, *Maple Transactions*, Vol. 3, No. 4, Article 17132, <https://doi.org/10.5206/mt.v3i4.17132>.

- [16] Teresa Gomez-Diaz, Tomas Recio (2024). The Conundrum Challenges for Research Software in Open Science, *Computers* 2024, 13(11), 302, <https://doi.org/10.3390/computers13110302>.
- [17] Teresa Gomez-Diaz, Tomas Recio (2026). Research software : qui es-tu ? Une conversation scientifique et juridique sur les logiciels de la recherche, 1024 - Bulletin de la société informatique de France, N. 27, <https://doi.org/10.48556/SIF.1024.27.53>.
- [18] Jean-Claude Guéron (2017). Open access: Toward the internet of the mind, The Budapest Open Access Initiative 15, <https://www.budapestopenaccessinitiative.org/boai15/>, https://openaccessprod.wpengine.com/wp-content/uploads/Guedon_BOAI15_FINAL.pdf.
- [19] Tiago Heinrich, Diego Kreutz, Roben Lunardi, Rodrigo Mansilha, Vinicius Fulber-Garcia, Lourenco Alves Pereira, and Rafael R. Obelheiro (2026). An Experience Report on Artifact Evaluation in Brazilian Conferences. *ACM SIGCOMM Computer Communication Review*, 56, 1, pp. 11-17, <https://doi.org/10.1145/3806097.3806100>.
- [20] Ben Hermann, Stefan Winter, Janet Siegmund (2020). Community Expectations for Research Artifacts and Evaluation Processes. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE'20)*, <https://doi.org/10.1145/3368089.3409767>.
- [21] Ben Hermann (2022). What Has Artifact Evaluation Ever Done for Us? *IEEE Security & Privacy*, 20(5), pp. 96-99, <https://doi.org/10.1109/MSEC.2022.3184234>.
- [22] Robert Heumüller, Sebastian Nielebock, Jacob Krüger & Frank Ortmeier (2020). Publish or perish, but do not forget your software artifacts, *Empirical Software Engineering* 25, pp. 4585-4616, <https://doi.org/10.1007/s10664-020-09851-6>.
- [23] Daniel Méndez Fernández, Wolfgang Böhm, Andreas Vogelsang, Jakob Mund, Manfred Broy, Marco Kuhrmann, Thorsten Weyer (2019). Artefacts in software engineering: a fundamental positioning, *Software & Systems Modeling*, 18, pp. 2777-2786, <https://doi.org/10.1007/s10270-019-00714-3>.
- [24] Daniel Mendez, Daniel Graziotin, Stefan Wagner, and Heidi Seibold (2020). Open science in software engineering. In: M. Felderer, G. Travassos (eds), *Contemporary Empirical Methods in Software Engineering*. Springer, Cham, pp. 477-501, https://doi.org/10.1007/978-3-030-32489-6_17.
- [25] National Information Standards Organization (NISO) (2021). NISO RP-31-2021, Reproducibility Badging and Definitions, Recommended Practice, <https://doi.org/10.3789/niso-rp-31-2021>.
- [26] Neeraj Shandilya, Marie-Sophie Barreau, Blanca Suarez-Merino, Andrea Porcari, Daniela Pimponi, Keld Alstrup Jensen, Wouter Fransman, Remy Franken (2023). TRAAC framework to improve regulatory acceptance and wider usability of tools and methods for safe innovation and sustainability of manufactured nanomaterials, *NanoImpact*, Volume 30, 100461, <https://doi.org/10.1016/j.impact.2023.100461>.
- [27] Christopher S. Timperley, Lauren Herckis, Claire Le Goues, Michael Hilton (2021). Understanding and improving artifact sharing in software engineering research, *Empirical Software Engineering* 26, 67, <http://dx.doi.org/10.1007/s10664-021-09973-5>.
- [28] Hugo Villamizar, Jannik Fischbach, Alexander Korn, Andreas Vogelsang, Daniel Mendez (2025). Prompts as Software Engineering Artifacts: A Research Agenda and Preliminary Findings. In: G. Scanniello, V. Lenarduzzi, S. Romano, S. Vegas, R. Francese (eds), *Product-Focused Software Process Improvement, PROFES 2025. Lecture Notes in Computer Science*, vol 16361, https://doi.org/10.1007/978-3-032-12089-2_32, <https://arxiv.org/abs/2509.17548>.
- [29] Stefan Winter, Christopher S. Timperley, Ben Hermann, Jürgen Cito, Jonathan Bell, Michael Hilton, Dirk Beyer (2022). A retrospective study of one decade of artifact evaluations, *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2022)*, pp. 145-156, <https://doi.org/10.1145/3540250.3549172>.