Towards a Framework Based on Open Science Practices for Promoting Reproducibility of Software Engineering Controlled Experiments

André F. R. Cordeiro [1]

[1]Informatics Department
State University of Maringá – Maringá, PR, Brazil

cordeiroandrefelipe@gmail.com

Abstract. *Experimentation in Software Engineering has increased in the last decades as a way to provide evidence on theories and technologies. In a controlled experiment life cycle, several artifacts are used/reused and even produced. Such artifacts are mostly in the form of data, which should favor the reproducibility of such experiments. In this context, reproducibility can be defined as the ability to reproduce a study. Different benefits, such as methodology and data reuse, can be achieved from this ability. Despite the recognized benefits, several challenges have been faced by researchers regarding the experiments' reproducibility capability. To overcome them, we understand that Open Science practices, related to provenance, preservation, and curation, might aid in improving such a capability. Therefore, in this paper, we present the proposal for an open science-based Framework to deal with controlled experiment research artifacts towards making such experiments de facto reproducible. To do so, different models associated with open science practices are planned to be integrated into the Framework.*

1. Introduction

Experimental Software Engineering (ESE) studies and investigates practices that can be applied in experiments carried out in Software Engineering (SE) [Wohlin et al. 2012, Shull et al. 2007]. Experiments can be represented from processes and stages. A process is presented in [Wohlin et al. 2012], with the stages of scoping, planning, operating, analyzing and interpreting, presenting and packaging, and recording. Figure 1 illustrates this process.

In Figure 1, it is possible to observe all stages of the experimental process. In the definition of scope, the experiment is evaluated, considering the problem under investigation and objectives. In planning, the context, hypotheses, variables, participating subjects (when applicable), experimental design, instrumentation, and assessment of threats to validity are defined. In the operation, data preparation, execution, and validation take place. During analysis and interpretation, data collected in the experiment are evaluated and hypothesis tests are performed.

In presenting and packaging, artifacts are organized for publishing and sharing. An artifact can be defined as every result used or generated during the experiment [Standard 2017]. Different artifacts are generated or manipulated during the experimental life cycle [Wohlin et al. 2012]. In the registration stage, the exposition of the results of

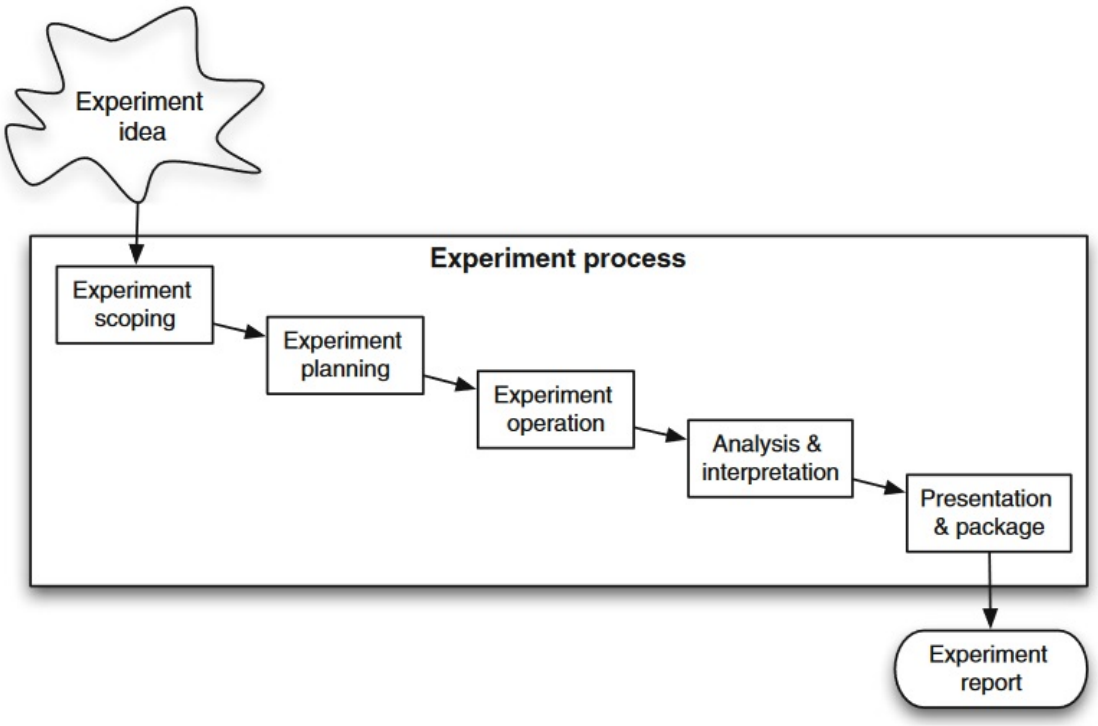

Figure 1. Experiment process [Wohlin et al. 2012].

the experiment is considered. Different means of dissemination can be considered, such as scientific articles and technical reports.

In addition to the experimental process, other subjects are investigated in ESE. Possible subjects include artifact management [Krishnamurthi 2013, Basili et al. 2007] and reproducibility [Li 2021, Liu et al. 2021, González-Barahona and Robles 2012]. Reproducibility is one of the basic rules of the scientific method and can be defined as the ability to reproduce a study, in full or in part, by different researchers [González-Barahona and Robles 2012]. The reuse of research artifacts and the advancement of scientific knowledge are examples of benefits directly related to reproducibility [Mendez et al. 2020, González-Barahona and Robles 2012].

Despite the importance of reproducibility for scientific advancement, different studies have reported the existence of reproducibility problems in SE, under different contexts, such as in Deep Learning (DL) models [Liu et al. 2021], secondary studies [Li 2021] and families of experiments [Kitchenham et al. 2020]. The sharing of artifacts used in experiments [Liu et al. 2021], the difficulty in adopting tools and practices in experiments [Anchundia et al. 2020], and the adoption of experimental packages [Shull et al. 2004] are examples of reported problems.

Problems observed in the reproduction of experiments carried out in SE motivated the proposal of a Framework. A Framework can be defined as a structure that can be refined and extended to support a set of functionalities [Standard 2017]. In the context of this study, a Framework is composed of data, metadata, and practices, focusing on reproducibility. The hypothesis considered is that data sets and metadata can favor reproducibility. Practices associated with Open Science (OS) can be considered in the

development of the Framework. OS can be defined as a movement that defends the reuse and sharing of all artifacts produced in scientific research [Mendez et al. 2020]. Different OS subareas have been established. A taxonomy with the main ones is presented in [Pontika et al. 2015]. Among the main subareas is Open Data (OD), defined as the ability to use, share and distribute one or more data sets, free of charge and with respect to defined licenses [Enríquez-Reyes et al. 2021, Osorio-Sanabria et al. 2020, Immonen et al. 2018, Pontika et al. 2015].

The next sections of this study describe the goals and research questions (Section 2), the problem domain (Section 3), the methodology (Section 4), the proposal of solution (Section 5), the research agenda (Section 6) and the final remarks (Section 7).

2. Goals and Research Questions

The objective of this study is to develop a Framework that favors the reproducibility of experiments carried out in SE. The Framework is based on data sets, metadata, and OS practices. In this context, the following research questions were established:

- Which data and metadata can be considered by the Framework?
- What practices associated with OS can be considered by the Framework?
- Is reproducibility favored when different data sets, metadata, and practices associated with OS about the experiment are available?

3. Problem Domain

During the planning, operation, and presentation of experiments, different data sets are constructed. Examples of data sets associated with the experimental process can be seen at [Wohlin et al. 2012]. Considering the knowledge of the experimental process and the problem of reproducibility (Section 1), it is observed that the experimental data sets have not been sufficient to favor the reproducibility of experiments carried out in SE. Difficulties observed with the adoption of experimental packages represent an example of this scenario [Shull et al. 2004].

The issue of reproducibility in SE can be studied and investigated from different perspectives. A study carried out with the aim of identifying tools that maximize reproducibility in SE experiments is presented in [Anchundia et al. 2020]. In [Liu et al. 2021], an investigation is presented on aspects of reproducibility and replicability in SE studies that consider the application of DL techniques. Problems with the reproducibility of secondary studies in Evidence-Based Software Engineering (EBSE) are presented in [Li 2021]. Investigations on the identification of families of experiments that used meta-analysis and evaluation of the reproducibility and validity of the results are presented in [Kitchenham et al. 2020].

In order to improve experimental replications, specific guidelines are presented in [Carver 2010]. The guidelines are proposed and organized into sets. To elaborate on these guidelines, different published replications were reviewed. An assessment, associated with reproducibility in software development, is presented in [Anda et al. 2008]. Among the results of the study is the finding that reproducibility in SE remains a major challenge, observed in research, education, and industry.

An initiative based on collaborative effort is proposed in [Neto et al. 2015], to favor the reproducibility of experiments related to the evaluation of software testing techniques. The development of a tool that performs the execution and analysis of experiments is also proposed. This structure was developed with the objective of helping in the creation and reproduction of experiments. The validation of the structure was carried out from the reproduction of a known experiment, with test case selection techniques. Regarding the reproducibility of studies in the software repositories mining area, elements that can be considered in the reproducibility are presented in [González-Barahona and Robles 2012]. Types of reproduction studies are presented, as well as a methodology to assess the reproducibility of studies.

4. Methodology

Initially, two secondary studies were considered. The first study is related to reproducibility in SE. The aim is to understand how reproducibility is represented and investigated. Some of the retrieved studies were described in Section 3.

The second study is related to the use of practices possibly associated with OS in SE. The objective is to understand which practices are used. Among the practices observed so far are OD offered as a software portal platform; definition of the open data model in applications; suggestions for developing open architectures and open software platforms; use of OD portals for software development. In this case, a portal is composed of different data sets; the use of OD clouds that consider linked OD and ontology; suggestions for developing a plan for creating and sharing artifacts, and recommendations to establish a long-term strategy for artifact sharing and evaluation.

Both secondary studies are in their final stages. The knowledge gained from these studies will guide the improvement of the Framework. At first, the Framework to be built will be finalized after two stages of development. In the first stage, conceptual development will take place, which consists of defining data sets, metadata, and practices. Recommendations to be considered by users of the Framework should also be described. In the second stage, the Framework will be implemented as a tool.

After completion of secondary studies, the following steps will be considered: identification of all models that form the Framework; identification of data and metadata associated with all models; evaluation and selection of OS practices to be used in the Framework; development of guidelines for the use of each Framework model; conceptual Framework development; implementation of the Framework as a tool. At the end of the implementation, it is possible to start the evaluation.

To evaluate the Framework, the empirical strategies of Case Study and Survey are considered [Wohlin et al. 2012]. In the Case Study, a scenario will be presented to illustrate the use of the Framework, in terms of data management, metadata, and practices. The management objective should be to favor the reproducibility of experiments in SE. Study participants must have knowledge of ESE. The aim is to enable its use by SE researchers.

In the Survey, the Case Study participants must evaluate the Framework based on a set of questions. Questions should be related to the ease of use and feasibility of adopting the Framework. Specific methodologies can also be considered. This is the case, for

example, of Technology Acceptance Model (TAM) [Davis 1985]. Another possibility of evaluation may be to carry out controlled experiments. A possible experiment suggestion is presented in [Cordeiro 2022].

5. Proposal of Solution

Considering the information presented in section 3, it is understood that data sets and metadata could complement the experimental data. Metadata can be defined as data about data [Yuan et al. 2013]. Personal experiences accumulated in previous projects suggest that metadata can be used to record characteristics of experimental data, in order to favor reproducibility. In the proposed Framework, the characteristics of curation, provenance, preservation, and data management are considered. Curation is associated with the description of metadata that describes actions performed with experimental data [Cordeiro et al. 2022]. Provenance makes it possible to record metadata related to the origin of the data. For example, someone can record the data collection location for a set [Yuan et al. 2013]. Preservation considers the registration of metadata related to the safe and complete maintenance of the artifacts considered [Cordeiro and OliveiraJr 2021]. In data management, there is a concern with sharing and reuse [Cordeiro et al. 2022].

To deal with the reproducibility of experiments performed in SE, a Framework was proposed. Previous experiences with experiments motivated the initial development of the Framework, presented in Figure 2.

When analyzing Figure 2, different elements can be noticed, such as conceptual model, recommender systems, ontology, trusted repositories, and model types. The conceptual model, the ontology, and the recommendation system are related to structures developed for recording and using experimental data [Furtado et al. 2021, Vignando et al. 2020]. The model types are related to metadata (curation, data management, provenance, and preservation) and OS practices.

6. Research Agenda

As mentioned in Section 4, two secondary studies are ongoing. At the end of these studies, the following research agenda will be initiated:

- Definition of all data sets that will be considered. For this definition, different experimental structures can be considered [Furtado et al. 2021, Vignando et al. 2020, Jedlitschka et al. 2008];
- Definition of all metadata sets, related to curation, data management, provenance, and preservation;
- Definition of OS practices associated with the Framework;
- Development of the different models shown in Figure 2. For each model, data, metadata, and OS practices will be considered;
- Conceptual development of the Framework, which consists of refining the structure develop in the previous step;
- Implementation of the Framework as a tool;
- Planning, elaboration, and execution of the Case Study;
- Planning, elaboration, and execution of the Survey;
- Availability of the tool for use by SE researchers.

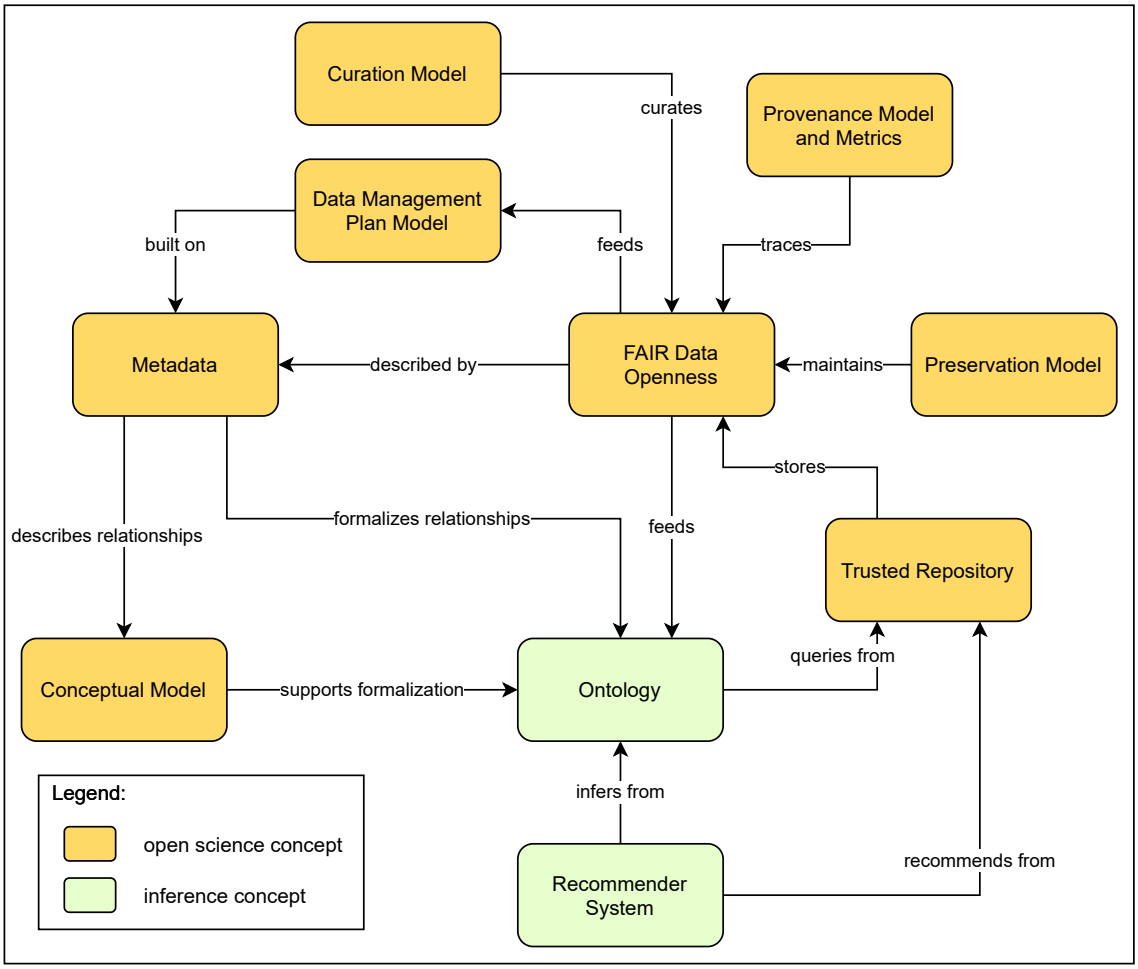

Figure 2. Initial Framework´s Structure [Cordeiro et al. 2022].

7. Final Remarks

This study proposed the development of a Framework to manage data and metadata of experiments performed with SE, focusing on reproducibility. Practices possibly associated with OS and being used in SE will be considered. The core element of the Framework is the data. Different sets of experimental data are from the beginning of the experimental process and represent the input to the Framework. In addition to these data sets, the definition of metadata sets is planned, and related to curation, preservation, provenance, and data management. Practices associated with OS should help organize and manage these sets. It is expected that the different sets of data and metadata can favor reproducibility.

The methodology for developing the research considers carrying out secondary studies, defining data sets and metadata, conceptual construction and implementation of the Framework, as well as carrying out a case study and a survey. In the end, the implemented Framework must be made available for use.

Acknowledgments

I would like to thank my advisor Prof. Dr. Edson OliveiraJr and my colleagues at the Research Group on Systematic Software Reuse and Continuous Experimentation (GREATER) for their support at different moments of the project.

References

Anchundia, C. E. et al. (2020). Resources for reproducibility of experiments in empirical software engineering: Topics derived from a secondary study. *IEEE Access*, 8:8992–9004.

Anda, B. C., Sjøberg, D. I., and Mockus, A. (2008). Variability and reproducibility in software engineering: A study of four companies that developed the same system. *IEEE Transactions on Software Engineering*, 35(3):407–429.

Basili, V. R., Zelkowitz, M. V., Sjøberg, D. I., Johnson, P., and Cowling, A. J. (2007). Protocols in the use of empirical software engineering artifacts. *Empirical Software Engineering*, 12(1):107–119.

Carver, J. C. (2010). Towards reporting guidelines for experimental replications: A proposal. In *1st international workshop on replication in empirical software engineering*, volume 1, pages 1–4.

Cordeiro, A. F. and OliveiraJr, E. (2021). Open science practices for software engineering controlled experiments and quasi-experiments. In *OpenScienSE*, pages 19–21, Brazil. SBC.

Cordeiro, A. F., OliveiraJr, E., and Capretz, L. (2022). Towards an open science-based framework for software engineering controlled (quasi-)experiments. In *Anais do XVI Brazilian e-Science Workshop*, pages 57–64, Porto Alegre, RS, Brasil. SBC.

Cordeiro, A. F. R. (2022). An open science-based framework for managing experimental data in software engineering. In *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022*, pages 342–346.

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. PhD thesis, Massachusetts Institute of Technology.

Enríquez-Reyes, R., Cadena-Vela, S., Fuster-Guilló, A., Mazón, J.-N., Ibáñez, L. D., and Simperl, E. (2021). Systematic mapping of open data studies: Classification and trends from a technological perspective. *IEEE Access*, 9:12968–12988.

Furtado, V., OliveiraJr, E., and Kalinowski, M. (2021). Guidelines for promoting software product line experiments. In *15th Brazilian Symposium on Software Components, Architectures, and Reuse*, pages 31–40.

González-Barahona, J. M. and Robles, G. (2012). On the reproducibility of empirical software engineering studies based on data retrieved from development repositories. *Empirical Software Engineering*, 17(1):75–89.

Immonen, A., Ovaska, E., and Paaso, T. (2018). Towards certified open data in digital service ecosystems. *Software Quality Journal*, 26(4):1257–1297.

Jedlitschka, A., Ciolkowski, M., and Pfahl, D. (2008). Reporting experiments in software engineering. In *Guide to advanced empirical software engineering*, pages 201–228. Springer, New York.

Kitchenham, B., Madeyski, L., and Brereton, P. (2020). Meta-analysis for families of experiments in software engineering: a systematic review and reproducibility and validity assessment. *Empirical Software Engineering*, 25:353–401.

Krishnamurthi, S. (2013). Artifact evaluation for software conferences. *ACM SIGSOFT Software Engineering Notes*, 38(3):7–10.

Li, Z. (2021). Stop building castles on a swamp! the crisis of reproducing automatic search in evidence-based software engineering. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pages 16–20. IEEE.

Liu, C., Gao, C., Xia, X., Lo, D., Grundy, J., and Yang, X. (2021). On the reproducibility and replicability of deep learning in software engineering. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(1):1–46.

Mendez, D., Graziotin, D., Wagner, S., and Seibold, H. (2020). Open science in software engineering. In *Contemporary empirical methods in software engineering*, pages 477–501. Springer.

Neto, F. G. D. O., Torkar, R., and Machado, P. D. (2015). An initiative to improve reproducibility and empirical evaluation of software testing techniques. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 2, pages 575–578. IEEE.

Osorio-Sanabria, M. A., Amaya-Fernández, F., and González-Zabala, M. (2020). Exploring the components of open data ecosystems: A systematic mapping study. In *Proceedings of the 10th Euro-American conference on telematics and information systems*, pages 1–6.

Pontika, N., Knoth, P., Cancellieri, M., and Pearce, S. (2015). Fostering open science to research using a taxonomy and an elearning portal. In *i-KNOW*, pages 1–8, New York. ACM.

Shull, F., Mendoncça, M. G., Basili, V., Carver, J., Maldonado, J. C., Fabbri, S., Travassos, G. H., and Ferreira, M. C. (2004). Knowledge-sharing issues in experimental software engineering. *Empirical Software Engineering*, 9:111–137.

Shull, F., Singer, J., and Sjøberg, D. I. (2007). *Guide to advanced empirical software engineering*. Springer, New York.

Standard, I. I. (2017). International standard - systems and software engineering – vocabulary. *ISO/IEC/IEEE 24765:2017(E)*, pages 1–541.

Vignando, H., Furtado, V. R., Teixeira, L. O., and OliveiraJr, E. (2020). Ontoexper-spl: An ontology for software product line experiments. In *ICEIS (2)*, pages 401–408.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media, Germany.

Yuan, D., Yang, Y., and Chen, J. (2013). 2 - literature review. In Yuan, D., Yang, Y., and Chen, J., editors, *Computation and Storage in the Cloud*, pages 5–13. Elsevier.