Towards an Open Science-Based Framework for Software Engineering Controlled (Quasi-)Experiments

André F. R. Cordeiro [1], Edson OliveiraJr[1], Luiz Fernando Capretz [2]

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

[2]Department of Electrical and Computer Engineering
Western University – London, Ontario, Canada

cordeiroandrefelipe@gmail.com, edson@din.uem.br, lcapretz@uwo.ca

Abstract. *Experimental Software Engineering has straightforwardly evolved in the last decades due to the effort of the community in providing consolidated training, teaching and practice. Particularly, for controlled experiments and quasi-experiments, the software engineering community has discussed on the lack of reproducibility and the missing of experimental artifacts sharing policies, such as, dataset, baselines, metamodels, repositories, and scripts. These are, therefore, important issues that jeopardizes controlled experimentation to evolve as rigorous as in millennial sciences as Medicine and Physics. In this ongoing work, it is presented a proposal of a conceptual framework for software engineering controlled experiments and quasi-experiments based on the main principles and practices of Open Science. It is understood that Open Science is one of the pillars to the evolution of science, consequently, to software engineering. The FAIR data, metadata, repositories, curation and provenance are some of the main practices discussed in this paper. Ongoing activities are described, in terms of how they are being performed and their relationship with prospective ones.*

1. Introduction

Experimental Software Engineering (ESE) investigates practices that can be adopted to improve experiments[1] performed in Software Engineering (SE) [Wohlin et al. 2012]. During the life cycle of an experiment, different stages should be considered, mainly focusing on planning, operation, and analysis and discussion of results [Wohlin et al. 2012].

Even with advances observed in the ESE area [OliveiraJr et al. 2021], there are still relevant issues to be addressed, such as, improving the level of reproducibility [González-Barahona and Robles 2012, Baker 2016, Anchundia et al. 2020, Nelson et al. 2021] and lack of sharing of experimental artifacts [Timperley et al. 2021, Damasceno et al. 2021], which prevent the software engineering evolution as a formal scientific discipline. It is understood that such an ESE evolution might be reached by increasing the formalization of the experiments carried out straightforwardly based on open science practices applied to SE [Mendez et al. 2020].

[1]In this paper when we refer to "experiment" we mean "experiments" and/or "quasi-experiments".

To do so, in previous works, our research group has carried out research on experiments formalization by: providing a panorama on how experiments are performed in SE; creating a set of guidelines to proper documenting SE experiments [Furtado et al. 2021]; and specifying a conceptual model to support the creation of an ontology for SE experiments [Vignando et al. 2020]. In recent studies, it was observed the possibility to incorporate practices related to the Open Science (OS)[2] context. OS can be defined as a scientific movement that stimulates the sharing of artifacts produced in scientific research to every citizen. Different subareas are related to OS [Medicine and others 2018], such as, open data, which has received increasing attention from the scientific community[3]. Recent works have demonstrated different actions related to the production, manipulation and availability of open data [Cordasco et al. 2018, Santos et al. 2018, Karanastasis et al. 2014], especially those related to FAIR data principles[4].

Based on the use of open data in Software Engineering, our hypothesis is that theinvestigation of solutions to gathering up and tracing data sets for SE experimentation might be a path to increase its reproducibility capacity and openness of ESE. Therefore, this paper presents the proposal of a framework based on OS practices for managing data related to controlled SE experiments taking advantage of OS practices. In this paper, the section 2 presents essential background on experimental software engineering and open science; Section 3 provides our first version of the proposed framework; Section 4 discusses how such a framework is being developed in terms of data provenance and data curation, as well as prospective actions; and Section 5 presents final remarks.

2. Theoretical Concepts

During an experiment different activities are performed [Wohlin et al. 2012]. The experimental process starts with the definition of the **scope** of the experiment with one or more objectives. In **planning**, the experiment is structured in terms of hypothesis formulation, variable selection, participant selection, selection of experimental design, instrumentation, and evaluation of validity. After planning, the **operation** takes place to collect data, which is **analyzed and interpreted** to draw conclusions regarding the established hypotheses. Next, the experimental packing is carried out to organize the artifacts for sharing. The sharing of these artifacts can aid in the experiment's reproducibility [González-Barahona and Robles 2012, Baker 2016, Anchundia et al. 2020, Nelson et al. 2021]. Finally, the experiment is reported [Wohlin et al. 2012].

During the experimental process, different artifacts are developed, used or reused, which represent results achieved during the experimental process [Jedlitschka et al. 2008]. Among possible artifacts are protocols, scripts and data [Mendez et al. 2020]. However, despite the body of knowledge already built in the ESE area, there are challenges to be overcome [Felderer and Travassos 2020, Wohlin et al. 2012, Shull et al. 2007], such as, those related to reproducibility [González-Barahona and Robles 2012], peer review [Ernst et al. 2021] and artifact sharing [Timperley et al. 2021, Damasceno et al. 2021].

[2]https://en.unesco.org/science-sustainable-future/open-science/recommendation

[3]https://www.fosteropenscience.eu/taxonomy/term/6

[4]https://www.go-fair.org/fair-principles

To deal with such challenges, it is suggested that the adoption of Open Science (OS) practices [5] might be a fair path. Open Science can be defined as a movement that considers the sharing of scientific artifacts for every citizen [Medicine and others 2018, Mendez et al. 2020]. In the context of OS, different subareas can be investigated[6] In [Pontika et al. 2015], the FOSTER open science taxonomy is described. This taxonomy presents some important subareas, as Open Access, Open Data, Open Reproducible Research, Assessment, Policies, and OS Tools. In the **open access** subarea, mechanisms are investigated to allow access to revised content, free of charge and with respected copyright. In the context of the **open data** subarea, there is an interest in the access of different data resources. In this way, data can be reused in other research and context. In **open reproducible research**, practices that may favor free access to experimental artifacts are investigated, to favor scientific reproduction. In the **evaluation of OS**, available results can be evaluated by the entire scientific community. In the **OS policy** subarea, guidelines are investigated for the application of practices at different levels, from groups to research institutions. In the case of **OS tools**, solutions are developed or investigated to aid the application of OS practices.

Some of these subareas have received more attention from the scientific community. Open data is an example. Recent papers in the literature have presented the importance of open data in different contexts [Cordasco et al. 2018, Santos et al. 2018, Karanastasis et al. 2014]. The framework described in this article considers two subareas, Open Data and Open Reproducible Research. All framework´s elements consider different data sets, important to facilitate the artifacts sharing, related to experiments. An expected result in this case is the higher capability to reproduce experiments in SE.

Details about the application of concepts related with Open Data and Open Reproducible Research, in framework´s context, are presented in next section. The framework presented in this article considers the data as central element. According to author´s experiences, it is believed that systematic management of experimental data in SE experiments can contribute to solve or minimize important issues in the ESE area, such as its reproducibility.

3. Open Science Framework for ESE

Data surround the experimentation process and all supporting elements of such a process as conceptual models, ontology, and automated systems [Cordeiro and OliveiraJr 2021]. The framework presented in this paper was planned with the aim of managing such data in every experimental activities in software engineering. Data generated in the experiments are not restricted to those for verification of the established hypotheses. Different types of data should be took into account to describe the experimentation environment. To deal with these types, specific elements are being considered for our experimental framework. Figure 1 presents an initial organization of the framework's elements.

Figure 1 and initial framework details are presented in [Cordeiro and OliveiraJr 2021]. This paper presents more details about the framework elements, in comparison with the mentioned paper. It is important to explain that data is the central element of the framework. The framework development is planned

[5] https://en.unesco.org/science-sustainable-future/open-science/recommendation

[6] https://www.fosteropenscience.eu/foster-taxonomy/open-science

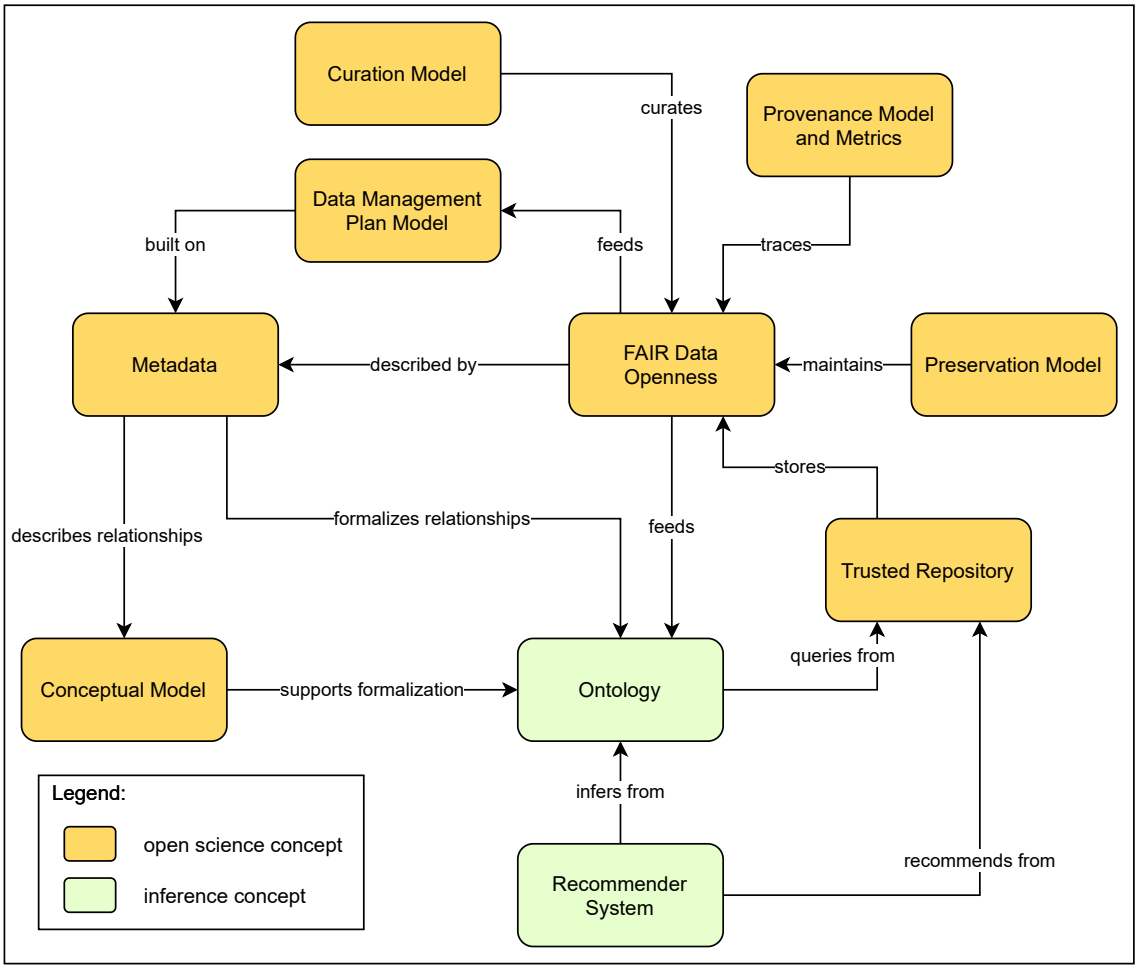

Figure 1. Initial framework organization [Cordeiro and OliveiraJr 2021].

at the level of specific open science practices elements as: FAIR data, preservation, provenance, curation, metadata and trusted repositories. In addition, conceptual model and ontology data models provide support to such a framework.

Data curation is related to activities to describe the data characteristics. Data provenance and metrics consider data´s source, traceability and future use. Data preservation concerns details on data's storage and deliverable capacity over time. Trusted repository focuses on how data is structured and reliably stored. Data management plan describes how experimental data and metadata are defined, their relationship and how data will be available for prospective interests. Metadata is related to important data attributes, in the context of FAIR data. Conceptual model describes essential experimental elements and their relationships to constitute the experimental process and created/reused artifacts. By developing specific models, one can deal with the characteristics of each element. For this paper, it is presented data provenance and data curation as practices to support our framework. The understanding of data provenance and curation context aids to comprehend the framework general purpose.

3.1. Data Provenance

Data provenance is a kind of important metadata in which the dependencies among application data sets are recorded [Yuan et al. 2013]. According to Yuan et al. "Data provenance records the information on how the data sets were generated, which is very important for our research on the trade-off between computation and storage". Data provenance has been constantly considered in scientific workflow systems.

The interest with data provenance can be observed in different contexts. In [Freund et al. 2019], the relationship between data provenance and information security is investigated. It is possible to verify that both areas benefit. Provenance is related to the record of people, institutions or entities that participated in manipulation of one or more data sets. Security is related to the search for confidentiality, integrity and availability of this data set. Costa et al. [Costa et al. 2019] describe an architecture for capturing and storing data about software development processes. A Provenance Model is considered by the architecture. Ontological model and data mining techniques are also considered, in order to identify opportunities for process improvement.

In addition to considering the provenance related to process activities, it is possible to consider artifacts generated in the execution of processes. In [Rousseau et al. 2020], an investigation is carried out to understand the possibilities for tracing the provenance of source code artifacts in a repository[7]. Still in the context of development, in [Tsai et al. 2007], a framework is proposed to analyze data from Service-Oriented Architectures (SOA) provenance. Provenance-related features in these systems are also analyzed. The experimental context is also considered at the provenance level. In [Alves et al. 2020], a taxonomy is presented to classify approaches that guide how to capture and store provenance data in simulation-based experiments in High Performance Processing (HPP) environments. In [Silva 2011], an architecture is presented for capturing and storing data generated in experiments carried out in computational clouds. The data to be captured and stored must favor the verifiability and reproducibility of the experiments.

3.2. Data Curation

Esteva et al. [Esteva et al. 2016] describe the curation performed on a data set, in the context of High Performance Computing (HPC). In the curation process, different activities were carried out, such as gathering requirements, definition of curation tasks, adding information elements, expanding the dataset scope, removing personal information, packaging the dataset into specific sizes and formats. Resende [Resende 2020] presents an investigation to understand the importance of digital curation activities for scientific data. The article explains the international trend of scientific knowledge management. In [Rocha and Gouveia 2020], curatorship is approached in the context of Education, in Higher Education Institutions, which work with Distance Education.

4. Discussion on Current Activities

The framework presented in this article is currently under development. The initial organization was shown in figure 1. Each element described in such figure will be addressed

[7]software heritage archive`https://www.softwareheritage.org/`

individually, then integrated to other elements. To do so, specific models will be developed and validated.

Currently, a systematic literature review is at late stage, aiming at gathering up and discussing existing OS practices in the software engineering area. From such a review, it will be developed a data provenance model for SE experiments. As the SE literature is scarce on this subject, it is possible that a novel data provenance model will be necessary. Such model will be based on experimental data and metadata already existing for SE experiments, developed in previous works [Vignando et al. 2020, Furtado et al. 2021].

Once a provenance model is developed, a curation model and policies/guidelines will be developed to allow experimental data to be incorporated to a new experiment, or even reused from previous studies. As curation policies will dictate how to handle such data to be available for current and prospective experiments, such a model will need to be capable of understand the provenance model to (semi-)automatize the (meta)data handling process. By handling experimental data and metadata, it will be take into consideration FAIR data lifecycle to allow data interchanging among experiments or even their several trials.

To build both provenance and curation models, it will be considered different supporting tools, such as, open science platforms and data/metadata manipulation tools. In addition to such an infrastructure, a web-based portal for supporting users of the framework will be developed as a front-end facility, aiming at increasing the adoption potential of the framework, such as the OSF[8] framework has been developed.

5. Final Remarks

This paper presented the incipient stages of development of open science-based framework for managing data generated or from SE controlled experiments and quasi-experiments. The framework is composed of different elements, which represent most know open science practices, such as, preservation, provenance, data management plan, metadata and trusted repositories. For each element, the development of a model is planned, which incorporates important data to the SE experimentation context.

At this moment, a systematic literature review is underway to understand the state of the art regarding open science practices for SE. As next steps it is necessary to conduct more advanced studies on each element, especially data provenance and curation. To do so, the development of specific models will be planned. Afterwards, empirical evaluation must be performed for each framework composing element. Integration is a must to allow all elements to work cooperatively.

References

Alves, R. C., Frota, Y., and de Oliveira, D. (2020). Gerência de dados de proveniência distribuídos de experimentos científicos: um mapeamento sistemático. In *Brazilian e-Science Workshop (BreSci)*, pages 97–104. in Portuguese.

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.

[8] `https://osf.io`

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604).

Cordasco, G., Malandrino, D., Pirozzi, D., Scarano, V., and Spagnuolo, C. (2018). A layered architecture for open data: Design, implementation and experiences. In *International Conference on Theory and Practice of Electronic Governance (ICEGOV)*, pages 371–381.

Cordeiro, A. F. and OliveiraJr, E. (2021). Open science practices for software engineering controlled experiments and quasi-experiments. In *Workshop de Práticas de Ciência Aberta para Engenharia de Software (OpenScienSE)*, pages 19–21.

Costa, G. C. B., Werner, C., Braga, R., Dalpra, H., Araújo, M. A., and Ströele, V. (2019). Deriving strategic information for software development processes using provenance data and ontology techniques. *International Journal of Business Process Integration and Management (IJBPIM)*, 9(3):170–196.

Damasceno, C., Melo, I., and Strüber, D. (2021). Towards multi-criteria prioritization of best practices in research artifact sharing. In *Workshop de Práticas de Ciência Aberta para Engenharia de Software (OpenScienSE)*, pages 1–6. Sociedade Brasileira de Computação (SBC).

Ernst, N. A., Carver, J. C., Mendez, D., and Torchiano, M. (2021). Understanding peer review of software engineering papers. *Empirical Software Engineering*, 26(5):1–29.

Esteva, M., Sweat, S., McLay, R., Xu, W., and Kulasekaran, S. (2016). Data curation with a focus on reuse. In *Joint Conference on Digital Libraries (JCDL)*, pages 45–54. IEEE.

Felderer, M. and Travassos, G. H. (2020). *Contemporary Empirical Methods in Software Engineering*. Springer.

Freund, G. P., Sembay, M. J., and Macedo, D. D. J. (2019). Data provenance and security of information: Interdisciplinary relations in the field of information science. *Revista Ibero-Americana de Ciência da Informação (RICI)*, 24(2):825–807.

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

Jedlitschka, A., Ciolkowski, M., and Pfahl, D. (2008). Reporting experiments in software engineering. In *Guide to Advanced Empirical Software Engineering*, pages 201–228. Springer.

Karanastasis, E., Andronikou, V., Chondrogiannis, E., Tsatsaronis, G., Eisinger, D., and Petrova, A. (2014). The opensciencelink architecture for novel services exploiting open access data in the biomedical domain. In *Panhellenic Conference on Informatics (PCI)*, pages 1–6.

Medicine and others, N. A. o. S. (2018). *Open Science by Design: Realizing a Vision for 21st Century Research*. National Academies Press.

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.

Nelson, N. C., Ichikawa, K., Chung, J., and Malik, M. M. (2021). Mapping the discursive dimensions of the reproducibility crisis: A mixed methods analysis. *PLOS One*, 16(7):e0254090.

OliveiraJr, E., Furtado, V., Vignando, H., Luz, C., Cordeiro, A., Steinmacher, I., and Zorzo, A. (2021). Towards improving experimentation in software engineering. In *Brazilian Symposium on Software Engineering (SBES)*, pages 335–340.

Pontika, N., Knoth, P., Cancellieri, M., and Pearce, S. (2015). Fostering open science to research using a taxonomy and an elearning portal. In *International Conference on Knowledge Technologies and Data-driven Business*, pages 1–8.

Resende, Lilian e Bax, M. (2020). Scientific data curation in information science: National scenario survey. *AtoZ: novas práticas em informação e conhecimento*, 9(1).

Rocha, D. G. and Gouveia, L. M. B. (2020). Digital content curation for distance education: Quality, updating and teaching skills. In *Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–4. IEEE.

Rousseau, G., Di Cosmo, R., and Zacchiroli, S. (2020). Software provenance tracking at the scale of public source code. *Empirical Software Engineering*, 25(4):2930–2959.

Santos, A. C., Pereira, Á. J., Oliveira, M. R., Macedo, H. T., and Nascimento, R. P. (2018). Building software products with use open data and big data in smart cities. In *Euro American Conference on Telematics and Information Systems (EATIS)*, pages 1–7.

Shull, F., Singer, J., and Sjøberg, D. I. (2007). *Guide to Advanced Empirical Software Engineering*. Springer.

Silva, C. (2011). Captura de dados de proveniência de workflows científicos em nuvens computacionais. Master's thesis, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa em Engenharia - Universidade Federal do Rio de Janeiro (COPPE/UFRJ). in Portuguese.

Timperley, C. S., Herckis, L., Le Goues, C., and Hilton, M. (2021). Understanding and improving artifact sharing in software engineering research. *Empirical Software Engineering*, 26:1–41.

Tsai, W.-T., Wei, X., Chen, Y., Paul, R., Chung, J.-Y., and Zhang, D. (2007). Data provenance in soa: Security, reliability, and integrity. *Service Oriented Computing and Applications*, 1(4):223–247.

Vignando, H., Furtado, V. R., Teixeira, L. O., and OliveiraJr, E. (2020). OntoExper-SPL: An ontology for software product line experiments. In *International Conference on Enterprise Information Systems (ICEIS)*, 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.

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.