Open Science Practices for Software Engineering Controlled Experiments and Quasi-Experiments[*]

André F. R. Cordeiro[1], Edson OliveiraJr[1]

[1]Informatics Department – State University of Maringá (UEM)
Maringá – PR – Brazil

cordeiroandrefelipe@gmail.com, edson@din.uem.br

Abstract. *In this vision paper, we present how Open Science practices can be adopted to prospectively support promoting software engineering controlled experiments and quasi-experiments. As experimentation in software engineering has gained extraordinary attention and increased in the last decade, we as community should focus on the openness of experiment artifacts and processes to every citizen, especially those artifacts produced with public and government funding. Such openness might bring several benefits towards evolving this area based on well-reported experiments, artifacts, processes, shared data, and experiences gathered up. In view of this, we envision an open science framework for software engineering controlled experiments ans quasi-experiments. In addition, we provide a research agenda, which is intended to be accomplished in the next five years.*

1. Context

During the experimental process, different artifacts are produced. Among them are protocols, scripts, and data [Wohlin et al. 2012]. The packaging and sharing of these artifacts facilitates and promotes the reproducibility of studies [Shull et al. 2007].

Although the literature presents a great extent of SE experiments[1] and certain replications, most of them do not provide a careful handling of their produced/used artifacts towards making them openly findable, accessible, interchangeable, and reusable (FAIR[2]). In addition, such artifacts are usually not treated to accomplish to the following Open Science principles: curation, provenance, preservation, or even trusted storage using repositories [Mendez et al. 2020, NASEM et al. 2018]. Such principles can contribute to Experimentation in Software Engineering (ESE), in terms of repeatability, replicability, and reproducibility [Mendez et al. 2020].

Among the main concepts presented by the Foster Open Science initiative and its taxonomy[3] are: open access, open data, open science evaluation, and open science policies. Open access considers free access to scientific materials, with little or no constraint. Open data provides free access to scientific data, for using, reusing and distribution. Open science evaluation is an open process for evaluating scientific results, not limited to peer

---

[1]Controlled experiments and quasi-experiments
[2]https://www.go-fair.org/fair-principles
[3]https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition

reviews. Open science policies considers recommendations for applying open science and achieving the associated objectives. Open science tools refer to a set of tools that can help in the open science development process. Considering these principles, we believe adopting Open Science can benefit ESE as we further describe.

2. Open Science for Software Engineering Experiments

We envision the adoption of different Open Science principles to SE experiments as a framework for this kind of study. Figure 1 depicts an initial draft of the framework.
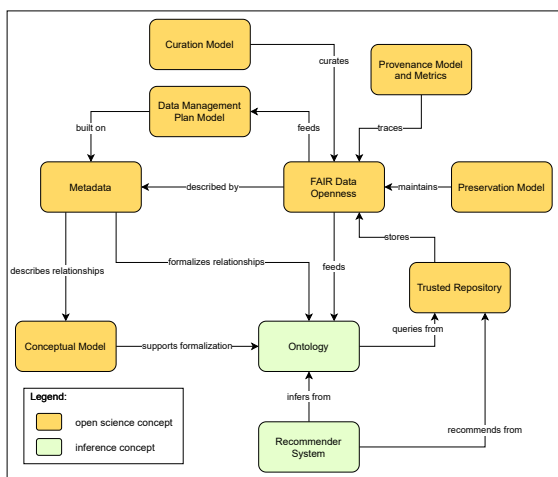
Figure 1. Open science practices framework for SE experiments

Every experiment produces several artifacts, but mainly data, which are our main entry point to start the openness process. With such data we envision making them FAIR, thus providing `FAIR Data Openness`, especially for data reuse and interchanging. However, such data requires `Metadata` to describe their structure, elements, and relationships. Metadata support experimental data as specified, for instance, in the Horizon 2020 worldly project guidelines[4]. With the experiment `Metadata`, we are able to build a `Data Management Plan Model`, which will describe how data acquired or produced during the experimentation process is managed, stored, what standards are used, and how they are handled and protected during and after an experiment. Then, `FAIR Data` feeds the Data Management Plan (DMP) model.

Acquired data might be from other experiments, from the same group, or reused from other groups. Therefore, we need to track inputs, systems, entities, and processes that influence the data of interest and provide historical records of the origin of them (e.g., in a database, document, repository, or even notebooks) with an explanation of how and why they got to the present experiment. Thus, we need to provide a `Provenance Model and Metrics` to keep track of such data.

Once data is acquired or produced they need to be kept safe to be reused and evolved in prospective reproducibility activities. Therefore, a `Preservation Model` provides a way to conserve and maintain the safety and integrity of such underlying data. Inconsistencies must be avoid by the aid of a metadata structure or even a meta-metadata level for these data quality attributes. Thus, data should never be lost or destroyed. In this context, `Trusted Repositories` are mandatory.

A `Trusted Open Repository` is an interesting alternative for the storage of experimental data and artifacts towards their preservation. The OpenAIRE initiative, for instance, tries to aid researchers at specifying trusted repositories by providing guidelines[5]

---

[4]`https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf`

[5]`https://www.openaire.eu/find-trustworthy-data-repository-certified-repositories`

to it, such as the ISO 16363[6] certification.

We understand a `Conceptual Model` is also needed to support the relationships of experimental concepts, especially those from `Metadata`. In addition, such a model should provide insights on how to formalize such concepts and relationships by building an `Ontology`, for example. An ontology is a way to group concepts on categories. Such experimental elements from the `Conceptual Model` have attributes, thus organizing them into classes and groups or clusters demands a formal definition of their relationships. `FAIR Data` is the source to feed the ontology, thus queries might be executed in the `Trusted Repository`.

Once an ontology is built, inferences might be performed over the ontology formalization, thus reflecting on how recommendations are provided based on the `Trusted Repository`. Therefore, recommendation on several different filters might done using a `Recommender System` to aid different kinds of users: students, instructors, practitioners, or even citizens in general.

3. An Agenda for Open Science Practices

Based on our vision on the openness of SE experiments, we propose an agenda for practicing open science principles within the next few years, as follows: (i) making experimental data complying with FAIR principles and establish common metadata to them; (ii) further investigating guidelines to build a trusted repository to store FAIR experimental data; (iii) definition of a conceptual model to aid establishing data and metadata relationships on experimental data; (iv) definition of curation, provenance and data management models for managing open experimental data and artifacts; (v) building of an ontology to support users inferring on experimental data and artifacts; and (vi) providing a mechanism to support recommending SE experiments for different users perspective.

We understand a prospective online supporting portal for guiding the framework users might be of great helpful and one of our products.

4. Final Remarks

We envisioned how open science principles and practices might support ESE data and artifacts. To do so, we provided a research agenda towards performing several activities to build an open science framework for SE experiments within the next few years.

At this proposal stage moment we envision certain extensions might be developed to other types of empirical studies such as for secondary studies protocols.

References

Mendez, D. et al. (2020). Open science in software engineering. In *Contemporary Empirical Methods in Software Engineering*, pages 477–501. Springer.

NASEM, N. A. o. S., Engineering, Medicine, et al. (2018). *Open science by design: Realizing a vision for 21st century research*. National Academies Press.

Shull, F. et al. (2007). *Guide to advanced empirical software engineering*. Springer.

Wohlin, C. et al. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

---

[6]https://www.iso.org/standard/56510.html