Specifying a Framework for Evaluating Requirements Engineering Technology

Challenges and Lessons Learned

Jose Luis de la Vara
Simula Research Laboratory
Lysaker, Norway
jdelavara@simula.no

Davide Falessi
Fraunhofer CESE
College Park, Maryland, USA
dfalessi@fc-md.umd.edu

Eric Verhulst
Altreonic NV
Linden, Belgium
eric.verhulst@altreonic.com

***Abstract*—Evaluating requirements engineering technology is a challenging activity. It becomes even more difficult when having to evaluate the technology and thus to show its suitability in real settings, as access to industrial resources might be limited and the target domain might be complex or very sensitive. This paper reports on our experience in specifying an evaluation framework for requirements engineering technology. The technology aims to improve safety assurance and certification practices, and is being developed in the scope of a large-scale European research project. We focus on presenting the challenges encountered and the lessons learned while specifying the framework. These lessons summarise how we addressed, plan to address, or propose to address the challenges. This information can be useful for other researchers and practitioners that have to evaluate requirements engineering technology in general, and with industry and for safety assurance and certification in particular.**

***Index Terms*—Evaluation, requirements engineering, evaluation framework, challenges, lessons learned, empirical software engineering, safety assurance, safety certification.**

I. Introduction

The importance of empirical evaluation of software engineering technology (e.g., tools, notations, techniques, and approaches [47]) has increasingly been acknowledged in the last years. Most of researchers would agree upon the need for providing empirical evidence of technology usability and usefulness [14] or for technology evaluation in industrial settings [4]. This is also the case in the requirements engineering (RE) community (e.g., [37][46]).

Evaluation of RE technology is never easy. It strongly depends on human aspects, thus on human involvement, and it is difficult to show how new technology can really improve industrial practices unless it is used in running projects [11]. In this paper we focus on the scenario in which RE technology has to be evaluated in industrial settings and with practitioners.

This type of evaluation can be especially challenging. Firstly, access to industrial resources (e.g., practitioners) can be limited. These resources can be expensive and not easy to obtain. The researchers must usually request them in advance and wait until they are available, and the importance of and the need for these resources for evaluation must be shown to those who have to provide the resources.

Secondly, evaluation in some domains is complex because of the inherent complexity of the domain. For example and in general, evaluating RE technology for avionics embedded systems can take much more time than for business information systems. Only understanding these systems is a challenge per se. It is unlikely that one person is an expert in all the possible domains and in their respective practices and systems.

Thirdly, some domains are very sensitive, not only to include data in publications, but also to provide data to researchers. A company can demand individual confidentiality agreements for the data accessed in a given moment. Strategies to mitigate this challenge must be defined and agreed upon with the companies involved in evaluation.

This paper reports on our experience in the specification of an evaluation for new technology targeted at safety assurance and certification. In essence, these are RE activities. They deal with systems whose safety requirements have to be analysed and met, whose safety has to be shown (i.e., it must be shown that the system can be deemed safe), and that must comply with safety regulations. The framework has been specified in the scope of OPENCOSS [33], a large-scale European research project that aims to devise a common certification framework for the automotive, avionics, and railway domains.

In addition to outlining the approach followed to specify the evaluation framework, we present the main challenges faced and the lessons learned from addressing them. The contribution of this information is two-fold. Firstly, the insights provided can help other researchers when having to specify evaluation frameworks, especially for evaluation in and with industry. Secondly, practitioners can benefit by increasing their awareness of the needs and challenges for their involvement in evaluating RE technology and for showing its benefits to them.

In addition, we think that this paper contributes to the maturity of RE research in general and of RE evaluation in real settings in particular. The latter is one of the weaknesses that most easily can be found in the literature (e.g., [18][31]). The paper also contributes to the development of a body of knowledge about how to perform these evaluations and how to address the challenges that can arise.

The rest of the paper is organized as follows. Section II presents the background of the paper. Section III outlines the approach followed to specify the evaluation framework. Section IV describes the challenges encountered and the lessons learned. Finally, Section V presents our conclusions.

II. Background

This section introduces the OPENCOSS project and reviews related work.

A. OPENCOSS

OPENCOSS is a FP7 European project that aims to (1) devise a common certification framework that spans different vertical markets for automotive, avionics, and railway industries, and (2) establish an open-source safety certification infrastructure. The ultimate goal of the project is to bring about substantial reductions in recurring safety certification costs and at the same time reduce certification risks through the introduction of more systematic safety assurance practices. The project deals with: (1) creation of a common certification conceptual framework; (2) compositional certification; (3) evolutionary chain of evidence; (4) transparent certification process, and; (5) compliance-aware development process. The project consortium consists of 17 partners from nine countries, and only four partners are from academia.

The task that corresponds to the experience presented in this paper aims to specify an evaluation framework and quality metrics for OPENCOSS results. These results are mainly a conceptual framework and a software tool, and will be evaluated in three cases studies [34]: development of a park system for an electric vehicle in the automotive domain, reuse of a railway execution platform (computing unit and operating system) in the avionics domain, and certification of a signalling system in the railway domain.

B. Related Work

This section reviews past work on evaluation of technology in industry, including RE technology. More insights into works related to safety assurance and certification are provided below. Review of existing metrics and frameworks was one of the activities that we performed for specifying the framework and in which we encountered some issues.

In general, the closest works to this paper are those that have reported case studies in which practitioners used, in real settings, the RE technology developed by the publication authors (e.g., [40]). This kind of studies have been identified in, for instance, literature reviews on RE such as [18]. In the area of safety assurance and certification, the ratio of these studies is low [29], especially if compared to other RE areas (e.g., traceability [31]).

Although the case studies reported in other publications and the insights provided are clearly valuable, the specific needs of OPENCOSS make its evaluation a more complex problem than the evaluations usually presented in the literature. For example, we aimed to specify a framework that fit three case studies and for which no much specific input was found. These are two of the challenges presented below.

Insights into how to deal with evaluation in industry, how to facilitate it, and how to promote technology transfer have been presented in works such as [5][17][49]. Some authors (e.g., [19]) have also provided guidance about how to conduct and report empirical evaluations in order to facilitate the analysis of its relevance and potential impact in industry.

Challenges for technology evaluation in industrial settings include providing short-term results, impacting practice, and understanding the domain. Some examples of suggestions to facilitate this evaluation are to have and show interest in solving industrial problems, to ask practitioners for early feedback on the solutions, and to find champions in industry.

Beyond supporting and presenting more evidence for the insights provided by other authors, this paper extends the state of the art by providing new insights (e.g., the difficulty of specifying a common evaluation framework for different domains) and concrete examples. The latter is a very important aspect for determining the extent to which the insights presented apply to other cases [19].

Finally, generic guidance for specifying evaluation frameworks can be found in works such as [2][21][41][42][48][50]. This guidance includes aspects such as checking existing frameworks, describing the phenomena under study in detail, formulating clear goals, and specifying procedures and guidelines for data collection and metric measurement.

III. Approach for Framework Specification

This section outlines the approach followed for specifying the evaluation framework. The approach was based on two main principles:

1. The comparison of the development and assurance processes of safety-critical system with and without using OPENCOSS results (i.e., its conceptual framework and software tool).
2. The use of the Goal-Question-Metric approach [2] to guide the design of the evaluation framework and the specification of its metrics.

The first principle aims to compare current practices with those enabled, facilitated, and/or supported by OPENCOSS results. Nonetheless, and as discussed below, we found some issues that made us change our vision about the possibility of the comparison.

The process followed for specifying the framework consisted of six main activities: (1) refinement of project goals and objectives, (2) agreement upon the empirical methods to use, (3) design, running, and analysis of a survey, (4) review of existing metrics and frameworks, (5) metrics specification, and (6) metrics tailoring for each case study in OPENCOSS. Most of the activities were executed in parallel and iteratively. For example, metrics were specified once the results of the survey were available and also as a result of the analysis of existing metrics and frameworks.

Furthermore, the results from other tasks were analysed. For example, we used the results of a systematic literature review [29] and of a survey on the state of the practice [30] concerning safety evidence management, and some requirements specification deliverables (e.g., regarding safety evidence management needs [35]).

The current version of the evaluation framework can be found in [36]. We plan to refine it when, for instance, more project results are available. The process followed is also inline with the guidelines provided in the works mentioned in the last paragraph of Section II.B. Indeed, we used them as references.

IV. Challenges and Lessons Learned

In this section we present the main challenges encountered when specifying the evaluation framework. Some are related to how the evaluation will have to be run. The importance of the challenges and the difficulty to address can be regarded as variable. Although some challenges might not be very difficult to tackle, all the challenges represent issues that we had to manage. All the lessons learned provide insights about how we did managed, plan to manage, or propose to manage the issues.

For each challenge, we describe it and explain the lessons learned from addressing the challenge. Such lessons learned provide suggestions for others about how similar situations can be addressed when having to evaluate RE technology. Although these challenges and lessons learned are closely related to OPENCOSS, we consider that they can arise in the evaluation of other technologies.

The challenges and the lessons learned are presented in the next subsections. The first six challenges are in the scope of the main activities for specifying the evaluation framework (Section III), whereas the rest are transversal.

C1: Refinement of Project Goals and Objectives

Description

The first activity towards specifying the evaluation framework was the analysis and refinement of the goals of OPENCOSS. Such goals had been specified in the project proposal, and were finally specified for the evaluation framework as follows:

- G1) To demonstrate a potential reduction of recurring costs for component/product safety certification across systems by 40% and across vertical markets by 30%
- G2) To demonstrate a potential reduction of certification risks by 20%
- G3) To demonstrate a potential gain for product innovation and upgrade by 20%

In relation to G2, we determined that certification risks [1] corresponded to: (1) the risk of not being able to create a system that can be deemed safe; (2) the risk of not being able to show that a system can be deemed safe, and; (3) the risk that a specific assessor or regulator will not agree upon deeming a system as safe.

In addition to refining the goals, we refined the objectives of the project for two main reasons. Firstly, we aimed to define more precise objectives. Secondly, the objectives had evolved as a result of the new insights gained and the new needs discovered since the project had started.

Lessons Learned

LL1: Evolving goals and objectives. The goals and objectives initially specified for a (research) project need to be analysed and might have to be refined when specifying its evaluation framework. It is likely that they have changed since they were specified.

LL2: Effect of new insights into project needs. If the definition of the problems to address and thus of the needs in a project change, then this change must be reflected in its evaluation framework.

C2: Agreement upon the Empirical Methods to Use

Although the evaluation had initially been envisioned as cases studies in which practitioners would use OPENCOSS results in real-life contexts and would compare the benefits of their use with their current practices, we soon realised that this might not be possible for all the results. Other evaluations might be more suitable. For example, some OPENCOSS results will corresponds to new practices in an organization, for which no past experience or data will be available.

As a consequence, OPENCOSS partners agreed that other empirical methods would also be necessary for evaluating project results. These methods are:

- Experiments, in which we will compare the results of executing some tasks with and without OPENCOSS results. For example, we plan to compare the gains in detecting errors in safety assurance documentation.
- Surveys, in which we will ask practitioners about their opinion regarding the use and possible benefits of OPENCOSS results. For example, we plan to ask for feedback on to what extent the use of the results can pay off, in relation to the necessary training and/or the requirements for their use.

We also realised that the benefits of some results could only be really evaluated in real, running projects, not only with some parts of them or with data from past projects. This is further discussed below.

We will also conduct lab demos and field trials [47] for initial evaluation. Lab demos will allow researchers to analyse OPENCOSS results with real data but not in the field, whereas field trials will allow them to analyse the results with real data and in the field (e.g., using the results with practitioners). In both cases, we will ask practitioners for feedback. Therefore, we will initially evaluate parts of the results (e.g., [12]) in reduced-scope settings in order to determine, for instance, their usability and potential usefulness.

Lessons Learned

LL3: Decision upon empirical methods. The foreseen need and suitability for an empirical method might change as further insights are gained into evaluation requirements.

LL4: Agreement upon empirical methods. It is necessary to agree upon the empirical methods to use for evaluating new technology. The agreement is essential with those providing evaluation data and with those having to participate in the evaluation.

LL5: Definition of an initial evaluation. Beyond the overall, ultimate evaluation of a new technology, it is important to evaluate parts of the technology in reduced-scope settings.

LL6: Participation of practitioners in all the evaluation stages. Practitioners should be involved in evaluation from its very beginning to its end, providing feedback as early as possible, regardless the scope of a specific evaluation activity.

C3: Need for a Survey

We conducted a survey among OPENCOSS partners in order to get input for the specification of the evaluation framework. As further discussed below, we had little input from the literature.

The survey focused on the effectiveness, efficiency, predictability, and scalability of safety assurance and certification processes, as the main quality aspects of the processes. For each aspect, several metrics were specified, and OPENCOSS partners were asked about these metrics.

Three OPENCOSS partners completed the survey. The main conclusions were as follows:

- All the quality aspects seem to be approximately equally difficult to meet.
- The results suggest that effectiveness is the most important quality attribute.
- For measuring effectiveness, error removal effectiveness, number of reused certified components, and proportion of successful deadlines seem to be the most suitable metrics.
- Stability of the process is the only metric that might not be very suitable for measuring efficiency. However, it is probably the most suitable metric for measuring predictability.
- Measurement of effort, error removal effectiveness, error removal efficiency, number of reused pieces of evidence, number of reused arguments, and number of reuse certified components can be largely supported by software tools, especially by Excel.
- The most costly to measure metrics seem to be the number of reused evidences, the number of reused arguments, and the number of certified components.
- Measurement of number of reused evidence and measurement of number of reused certified component seem to be the metrics that require more automated support.

Details about the research questions, the questionnaire, and the results of the survey can be found in [36].

Lessons Learned

LL7: Suitability of a survey. A survey is an easy way to obtain input from practitioners for an evaluation framework.

LL8: Importance of a survey. No much information might be available about practitioners' perspectives and opinions regarding the importance and need of measuring some aspects in an evaluation framework. A survey is an excellent method to obtain such perspectives and opinions, and can turn to be essential for the success of the framework.

C4: Insufficient Input from Existing Metrics and Frameworks

Although several works related to the evaluation framework and the quality metrics for OPENCOSS results were identified, their value as input was limited. The main reason was that the metrics provided could not be directly (re)used for evaluation of OPENCOSS results. We also tried to get input from more general areas, aiming to gain insights that could facilitate the definition of the evaluation framework.

Related work was divided into three categories: evaluation of safety assurance and certification, software and software process metrics, and safety metrics.

In relation to **evaluation of safety assurance and certification**, the results of a systematic review on safety evidence [29] showed that past research on safety certification had barely performed empirical studies, evaluated research results in industrial settings, and thus presented metrics applied in safety assurance and certification processes. This was in line with the acknowledged lack of well-defined, measurable safety assurance metrics [24].

Nonetheless, some studies had provided insights and defined metrics that could be useful for the evaluation framework. The areas studied by those research works were: software safety measurement and process improvement [3]; validation of safety-critical systems [8]; project management for safety assurance [9]; safety certification of airborne software [13]; use of COTS in safety-critical systems [22]; development process of safety-critical systems [25]; measurement of process risk [26]; safety case patterns [44], and; software reuse in systems of systems [45]. Some tools for safety case development and evidence management had also addressed definition and management of metrics [7][10][15].

We also checked the deliverables of some related projects (e.g., [6][39]). Although they provided some insights into the evaluation of their results, these deliverables did not specify detailed, quantitative metrics, but only questions or aspects for measurement of project goals.

Software and software process metrics is an area that has made an important progress for the last three decades. Nowadays, it can be regarded as a mature area. However, in OPENCOSS we are not directly managing software products or software processes, but focus on specific, special aspects such as safety assurance processes. This means that the past work on software and software process metrics did not directly meet our evaluation needs. Nonetheless, we considered that its overall ideas, principles, areas of study, and metrics could be used as input and reflected in the evaluation framework.

Well-known books on software and software process metrics can easily be found (e.g., [16]), and in recent years systematic literature reviews [43] and mapping studies [23] have also been conducted. Typical examples of indicators for software and/or software process measurement that can be found in these works are: complexity, cost, defects, effort, estimation accuracy, process quality, productivity, product quality, schedule, size, stability, and time-to market.

Our conclusions from searching for **safety metrics** were similar to those about software and software process metrics. Although not directly applicable, the ideas and areas of study could be used as input for the evaluation framework, especially in relation to G2. The metrics were not explicitly targeted at safety assurance and certification (e.g., the number of hazards of a system does not say much about safety assurance per se), but could be adapted for OPENCOSS evaluation purposes.

Most of the literature on safety metrics had discussed safety risk definition and measurement (e.g., [27]). We also found works on safety metrics related to managerial aspects (e.g., related to safety training [20]), software safety (e.g., [32]), safety performance (e.g., [38]), and process safety (e.g. [28]).

Lessons Learned

LL9: A lot of work on metrics is available. There are many works that have dealt with technology evaluation in the past and proposed metrics. Nonetheless, their suitability for a given evaluation framework must be analysed.

LL10: Lack of reference metrics and frameworks. It is possible that no much work has been performed in relation to the specification of evaluation frameworks and metrics in the scope of specific evaluation needs.

LL11: Usefulness of existing metrics and frameworks. Although no much specific input is available for an evaluation framework, past work can still provide useful insights and serve as a reference. The possibility of adapting it has to be determined.

C5: Metrics Specification

The activity with the highest degree of collaboration among the contributors to the evaluation framework was metrics specification. On the one hand, some partners (including ourselves) were responsible for specifying the questions for each project goal and subsequently the metrics for each question. A formula for calculation and a description was provided for each metric. On the other hand, all the partners provided feedback on the questions and the metrics. Firstly, they indicated if more aspects (i.e., questions) should be studied. Secondly, they indicated if they considered the metrics to be suitable and if their descriptions were understandable.

The main, overall aspects addressed in the questions were: (1) productivity; (2) rework needs; (3) number of residual defects; (4) defect density; (5) automatic element creation; (6) reuse; (7) time aspects; (8) workload/resources aspects, (9) defect removal costs and time; (10) efficiency; (11) cost; (12) difficulty of cross-certification, and; (13) awareness of necessary work.

Examples of the questions formulated for the project goals and of the metrics specified for each questions are as follows:

G1) Reduction of recurring cost

- How can the safety assurance process be efficient for long system lifecycle (delta demonstration)? (e.g., effort for determining the work required for product reuse)
- How can safety assurance be reused across systems? (e.g., certification requirements fulfilled)

G2) reduction of certification risks

- How can we gain early insights into the state of safety for certification purposes? (e.g., early risk detection)
- How does confidence in the safety assurance process relate to certification risks? (e.g., change requests by assessors)

G3) Gain for product innovation and upgrade

- How can safety demonstration be used in product upgrade? (e.g., safety assurance assets reused for product upgrade)
- How can product upgrade certification be cost-effective (delta certification)? (e.g., re-certification effectiveness)

We specified a set of 13 questions and a set of 41 metrics. More details about them are available in [36].

Lessons Learned

LL12: Request for feedback on metrics. It is important to ask evaluation stakeholders their opinion about the aspects to address and the metrics to measure in an evaluation framework.

LL13: Agreement upon the aspects to address and the metrics to measure. The aspects to address and the metrics to measure in an evaluation framework should be agreed upon with the evaluation stakeholders.

C6: Metrics Tailoring for each Case Study

Once the metrics had been specified and agreed upon, we ask the companies that provide the case studies to specify the procedures and guidelines to measure the metrics in the case studies. For each metric, and in relation to each case study, we determined if the metric: (1) was relevant; (2) was not directly and explicitly relevant, but it could be measured in the case study by adapting it; (3) was not relevant, or; (4) might be relevant, but this had to be studied in more depth in the future.

A description of its relevance to a case study and of how it could be measured was provided for each metric. An estimation of the improvement from the current situation to the one with OPENCOSS results was provided when possible.

Lessons Learned

LL14: Metric analysis for different case studies. When having to evaluate technology in several case studies, the relevance of and the possibility of measuring the metrics of an evaluation framework must be determined for each case study.

LL15: Request for information about metric measurement procedures and guidelines. Those providing case studies should be asked to prepare a description of how the metrics of an evaluation framework could be measured in their cases.

C7: Specification of a Common Evaluation Framework

Description

The initial plan was to provide a common evaluation framework for the three case studies in the three domains addressed. However, providing such a common framework turned to be a very difficult objective, maybe unrealistic.

Firstly, the state of the practice and the current situation in each application domain is not homogeneous. For example, compliance with safety standards is very recent in the automotive domain, and some aspects relevant for the avionics and railway domains are not considered for safety assurance and certification yet. Secondly, each case study is different and targeted at evaluating specific OPENCOSS results. The automotive case study deals with compositional certification, the avionics case study focuses on cross-domain certification, and the railway case study pays special attention to safety evidence traceability. Therefore, all the questions and metrics were not equally relevant or applicable to all the case studies.

In summary, the evaluation framework finally aims to provide a generic set of questions and metrics, which can be refined and adapted to the context of each case study. It is not possible to measure all the metrics in all the case studies because of their differences. Nonetheless, if considered necessary, it might be decided to adapt and extend the case studies for validation and evaluation purposes.

Lessons Learned

LL16: Unfeasibility of a multi-domain framework. Specifying a common, detailed evaluation framework for different domains and case studies might not be feasible.

LL17: General study of the phenomena to evaluate. When aiming to specify a common evaluation framework, the phenomena under study should be analysed from a generic, general perspective. The framework can later be refined and adapted to the aspects specific to a given case.

C8: Possibility of Evaluating some Aspects only in Real, Running Projects

Description

Enactment of the evaluation framework aims to show the benefits from using OPENCOSS results in current practices on safety assurance and certification. This will be mainly measured in the case studies of each application domain.

However, there are some practices that can benefit from OPENCOSS results but whose real impact will only be possible to measure accurately in real, running projects. Although representative scenarios of safety assurance and certification processes are enacted, they might not reflect the actual impact of using OPENCOSS results. For example, their real impact on cost reduction can only be estimated in running projects. As a consequence, and also related to the previous challenge, it can be determined that some metrics will not be measured during OPENCOSS or that will have to be adapted.

Lessons Learned

LL18: Identification of the metrics to measure in real projects. The metrics whose measurements are only relevant in real, running projects must be identified and agreed upon with the evaluation stakeholders.

LL19: Decision upon the metrics to measure in real projects. For those metrics than can only be measured in real, running projects, the companies that might adopt the technology have to decide if they can and want to measure the metrics in the timeframe of a (research) project.

C9: Lack of Information about the Current Situation

Description

Based on the initial information gathered to analyse the current practices for safety assurance and certification, we concluded that evaluating the benefits from OPENCOSS results might be more complex than anticipated. This is a result of the lack of details and of the lack of a common understanding for some aspects. For example:

- Many projects related to safety certification have a long lifecycle, in combination with incremental modifications. It is not trivial to extract cost and resource data that can be allocated to only a specific project.
- In the automotive domain, certification is not really an issue yet. For the pre-cursor (conformity assessment), no systematic records are kept as the data is spread over the tier 1-2-3 supply chain.
- Only one respondent of the survey was able to provide an estimation of the cost of a safety assurance and certification project.

Therefore, we drew the conclusion that some metrics should be based on a sampling of selected action points that are relevant in the context of developing a certifiable product. We also considered that some estimation should be provided for those metrics for which no accurate information about the current situation was available.

Lessons Learned

LL20: Estimation for metrics. If accurate values for some metrics cannot be provided regarding the current situation, then practitioners must at least try to provide an estimate.

LL21: Decision upon estimation of metrics. If no accurate values can be provided for a metric, then it must be decided if the metric should be changed or if an estimate can suffice.

C10: Need for More Knowledge about Project Results

Description

From a general perspective, the evaluation framework aims to analyse the achievement of OPENCOSS goals (G1, G2, and G3) based on the new possibilities that OPENCOSS results will enable for safety assurance and certification. Consequently, the framework directly depends on the conceptual and software solutions that the project provides.

At this stage of the project, many details of OPENCOSS and its results still have to be defined in more detail. For example, the current status of the requirements specified is "proposed", but we do not know for sure yet which requirements will finally be implemented. This means that a completely clear vision of what OPENCOSS will provide is not available yet, and thus that it is very difficult to determine the actual possibility and relevance of the evaluation of some results that OPENCOSS could provide.

In addition, some aspects of the project vision have evolved. The detailed analysis of the state of the practice and the new insights gained into it have allowed us to discover "hidden" issues that had not been taken into account in the project initially, have to be addressed in the near future in the project, and whose solutions will have to be evaluated.

As a result, the evaluation framework will evolve and be updated in the future, iteratively, as the work progresses and thus we more clearly understand what can be expected from the OPENCOSS conceptual framework and the OPENCOSS tool platform. As mentioned above, the applicability of some metrics in some case studies will be determined in the future.

Lessons Learned

LL22: Evaluation frameworks can evolve. In a research project, an evaluation framework corresponds to an artefact that will very likely evolve as the project does.

LL23: Evaluation frameworks might not need to be perfect from their first version. Especially in a research project, the evaluation stakeholders should not be concerned about having a perfect evaluation framework from the very beginning. The framework will have to be adapted as the project's vision and requirements evolve. The important point is to have a stable, well-defined framework when the evaluation starts.

LL24: Iterative approach for evaluation frameworks. For an "evolving" evaluation framework, it is important to try to agree with the rest of evaluation stakeholders upon an iterative approach for updating and refining the framework, determining what is expected for the next iteration.

C11: Confidentiality Issues

Description

Another issue encountered was that a complete data set of the case studies was not available for confidentiality and competitive pressure reasons. As mitigation measures, the following action lines were agreed upon:

- The industrial partners would sanitise the case study data.

- The scope of the evaluation would initially be narrowed to activities and data related to hazard and risk analysis and to software validation and verification.
- The industrial partners would use the internally available data and provide the consortium with the measured or assessed improvement figures. This way no sensitive data needs to be communicated.

Lessons Learned

LL25: Need for a strategy for data release. The strategy for releasing evaluation data to researchers should be agreed as soon as possible. It will help researchers to plan their work.

LL26: Use of data sub-sets. If it is not possible to obtain the entire information of a project, then the evaluation stakeholders should aim to find a suitable part of the project for evaluation.

LL27: Awareness of data release consequences. Companies owning the evaluation data must be (made) aware of the consequences of not providing data or evaluation results.

LL28: Consequences of restricted data release. If no data is provided for some evaluation activities, researchers' work might be hindered and the companies owning the data might have to analyse the results themselves, without any support from others.

C12: Unclear Relationship of Process Improvement with Better Practices and Product Quality

Description

An issue that complicates drawing clear conclusions about improved cost-efficiency for certification is that any metric is likely to be related to several aspects and factors simultaneously. Certification as such is usually an end-stage, after-development activity. Although it looks at quality and process aspects, these aspects are usually organisation and/or product family specific and wide.

We can certainly assume a positive correlation between (1) a more controlled process and (2) the capability to have a product certified and higher product quality. However, this does not exclude the fact that a lower quality product or a less controlled process can still result in a positive certification outcome. This also means that a product that is safety certified is not necessarily more reliable. Although related, both aspects are different.

Lessons Learned

LL29: Impossibility of determining cause-effect relationships. It might not be possible to find a cause-effect relationship for important aspects of a research project.

LL30: Importance of cause-effect that cannot be determined. Although it might not be possible to determine a clear cause-effect relationship for some variables of an evaluation framework, studying these variables can still provide important and valuable insights for others.

V. Conclusion

This paper has presented our experience in specifying an evaluation framework for requirements engineering technology. More concretely, the technology corresponds to a conceptual framework and tool support for safety assurance and certification in the context of a large-scale European research project. Three case studies from the automotive, avionics, and railway domains will be used for evaluation.

We faced several challenges while specifying the framework. Some challenges were in the scope of the main activities executed: refinement of project goals and objectives, agreement upon the empirical methods to use, design, running, and and analysis of a survey, review of existing metrics and frameworks, metric specification, and metric tailoring to each case study. The specification of a common evaluation framework, the possibility of evaluating some aspects only in real, running projects, the lack of information about the current situation, the need for more knowledge about project results, confidentiality issues, and the unclear relationship of process improvement with better practices and product quality were transversal challenges.

Facing these challenges allowed us to learn and thus to provide a set of 30 lessons. In our opinion, the main meta-lessons are the evolutionary nature of an evaluation framework and of the input for creating it, the importance of finding or creating suitable input, the need for communication and agreements among the evaluation stakeholders, and the specific issues in dealing with generic evaluation frameworks.

The lessons presented can be very valuable and useful for both academia and industry. They can help researchers and practitioners to better know the needs and problems of similar situations and to define strategies to address them. The insights can be especially relevant for projects in which industry and academia have to collaborate.

We have mentioned above some future work, such as refining the evaluation framework. This might lead to the discovery of new challenges and to learning new lessons. We also plan to define a catalogue of metrics for safety assurance and certification. We would like to conduct a systematic review on the topic, define metrics based on the project's conceptual results (e.g., [12]), and validate the metrics.

Acknowledgment

The research leading to this paper has received funding from the FP7 programme under the grant agreement n° 289011 (OPENCOSS) and from the Research Council of Norway under the project Certus SFI. We would also like to thank the OPENCOSS partners that have provided input and feedback on the evaluation framework, especially to those who contributed to its specification: Fabien Belmonte, Cedric Chevrel, Marc Fumey, Sandrine Georges, Vincenzo Manni, Alberto Melzi, Sunil Nair, Florent Pages, Laurent Pitot-de-la-Beaujardiere, Mehrdad Sabetzadeh, and Giorgio Tagliaferri.

References

[1] R. Alexander, T. Kelly, and B. Gorry, "Safety Lifecycle Activities for Autonomous Systems Development", in 5th SEAS DTC Technical Conference, 2010

[2] V.R. Basili, G. Caldiera, and H.D. Rombach, "The Goal Question Metrics Approach", in Encyclopaedia of Software Engineering, vol. I, 1st ed., pp. 528-532, Wiley (1994)

[3] V.R. Basili, et al., "Obtaining valid safety data for software safety measurement and process improvement", In ESEM 2010

[4] L.C. Briand, “Embracing the Engineering Side of Software Engineering”, IEEE Software, vol. 29(4), p. 96, 2012

[5] L.C. Briand, et al., “Research-Based Innovation: A Tale of Three Projects in Model-Driven Engineering”, in MoDELS 2012, pp. 793-809

[6] CESAR project, “Deliverable D_SP5_R4.4_M4 - Public evaluation report for Automotive RTP V3”, 2012

[7] CertWare, http://nasa.github.com/CertWare/

[8] K.J. Cruickshank, J.B. Michael, and M.T Shing, “A Validation Metrics Framework for Safety-Critical Software-Intensive Systems”, in SoSE 2009, pp. 1-8

[9] DACS, “Software Project Management for Software Assurance - A DACS State-of-the-Art Report”, Tech. Report 347617, 2007

[10] E. Denney, G. Pai, and J. Pohl, “Automating the Generation of Heterogeneous Aviation Safety Cases”, Technical Report, NASA/CR–2011–215983, 2011

[11] J.L. de la Vara, et al., “Towards Customer-Based Requirements Engineering Practices”, in EmpiRE 2012, pp. 37-40

[12] J.L. de la Vara, S. Nair, and R.K. Panesar-Walawege, “On the Use of Artefacts as Safety Evidence: A Conceptual Model”, Simula Research Laboratory, Technical Report, 2013

[13] I. Dodd and I. Habli, “Safety certification of airborne software: An empirical study”, Reliability Engineering & System Safety, vol. 98(1), pp. 7-23, 2012

[14] T. Dybå, B.A. Kitchenham, and M. Jørgensen, “Evidence-Based Software Engineering for Practitioners”, IEEE Software, vol. 22(1), pp. 58-65, 2005

[15] D. Falessi, et al., “Planning for Safety Evidence Collection: A Tool-Supported Approach Based on Modeling of Standards Compliance Information”, IEEE Software vol. 29(3), pp. 64-70, 2012

[16] N.E. Fenton and S.L. Pfleeger, Software Metrics - A Rigorous & Practical Approach, 2nd ed., PWS 1998

[17] T. Gorschek, et al., “A Model for Technology Transfer in Practice”, IEEE Software, vol. 23(6), pp. 88-95, 2006

[18] M. Ivarsson and T. Gorschek, “Technology transfer decision support in requirements engineering research: a systematic review of REj”, Requirements Engineering, vol. 14(3), pp. 155-175, 2009

[19] M. Ivarsson and T. Gorschek, “A method for evaluating rigor and industrial relevance of technology evaluations”, Empirical Software Engineering Journal, vol. 16(3), pp. 365-395, 2011

[20] C.A. Janicak, Safety Metrics: Tools and Techniques for Measuring Safety Performance, Government Institutes, 2003

[21] N. Juristo and A.M. Moreno, Basics of Software Engineering Experimentation, Springer, 2001

[22] E. Kesseler, “Assessing COTS software in a certifiable safety-critical domain”, Information Systems Journal, vol. 18(3), pp. 299-324, 2008

[23] B. Kitchenham, “What’s up with software metrics? - A preliminary mapping study”, Journal of Systems and Software, vol. 83(1), pp. 37-51, 2010

[24] A.J. Kornecki and J. Zalewski, “Selected Issues in Computer Systems Safety: Position Paper”, in 1st IEEE International Workshop on Safety of Systems, 2007

[25] K. Kvinnesland, “Implementation of metrics in development of highly-safety critical software”, in EuroSPI98

[26] L. Layman, et al., “A case study of measuring process risk for early insights into software safety”, in ICSE 2011, pp. 623-6332

[27] R. Maguire, Safety Cases and Safety Reports - Meaning, Motivation and Management, Ashgate, 2006

[28] J. Murdoch, G. Clark, and A. Powell, “Measuring Safety: Applying PSM to the System Safety Domain”, in SCS’03, pp. 47-55

[29] S. Nair, et al., “Classification, Structuring, and Assessment of Evidence For Safety: A Systematic Literature Review”, in ICST 2013

[30] S. Nair, et al., “The State of the Practice on Evidence Management for Compliance with Safety Standards”, Simula Research Laboratory, Technical Report, 2013

[31] S. Nair, J.L. de la Vara, and S. Sen, “A Review of Traceability Research at the Requirements Engineering Conference”, in RE 2013 (accepted paper)

[32] NASA, “NASA Software Safety Guidebook”, NASA Technical Standard NASA-GB-8719.13, 2004

[33] OPENCOSS project, http://opencoss-project.eu

[34] OPENCOSS project, “Deliverable 1.2 - Use cases description and business impact”, 2012

[35] OPENCOSS project, “Deliverable D6.2 - Detailed requirements for evidence management of the OPENCOSS platform”, 2012

[36] OPENCOSS project, “Deliverable D1.3 - Evaluation framework and quality metrics”, 2013

[37] B. Paech, et al., “An Analysis of Empirical Requirements Engineering Survey Data”, in Engineering and Managing Software Requirements, pp. 427-452, Springer, 2005

[38] R. Pitblado, “Real-Time Safety Metrics and Risk-Based Operations”, in 11th Int. Symposium on Loss Prevention, 2004

[39] RECOMP project, “Deliverable D6.1 - Evaluation metrics definition”, 2012

[40] B. Regnell, R. Berntsson-Svensson, T. Olsson, “Supporting Roadmapping of Quality Requirements”, IEEE Software, vol. 25(2), 42-47, 2008

[41] P. Runeson, et al., Case Study Research in Software Engineering: Guidelines and Examples, Wiley, 2013

[42] F. Shull, J. Singer, and D.I.K. Sjøberg, (eds.), Guide to Advanced Empirical Software Engineering, Springer, 2008

[43] M. Unterkalmsteiner, et al., “Evaluation and Measurement of Software Process Improvement - A Systematic Literature Review”, IEEE Transactions on Software Engineering, vol. 38(2), pp. 398-424 2012

[44] S. Wagner, et al., “A Case Study on Safety Cases in the Automotive Domain”, in ISSRE 2010, pp. 269-278

[45] B. Warren, J.B. Michael, M.T. Shing, “A Framework for Software Reuse in Safety-Critical System-of- Systems”, in SoSE 2008, pp. 1-6

[46] R. Wieringa, “Requirements researchers: are we really doing research?”, Requirements Engineering Journal, vol. 10(4), pp. 304-306, 2005

[47] R. Wieringa, “Requirements Engineering Research Methodology: Principles and practice”, in RE 2008 Tutorials

[48] R. Wieringa, “Towards a unified checklist for empirical research in software engineering”, in EASE 2012, pp. 161-165

[49] C. Wohlin, et al., “The Success Factors Powering Industry-Academia Collaboration”, IEEE Software vol. 29(2), pp. 67-73, 2012

[50] C. Wohlin, et al., Experimentation in Software Engineering, 2nd ed., Springer, 2012