Towards a Semantic Knowledge Base on Threats to Validity and Control Actions in Controlled Experiments

Stefan Biffl [a], Marcos Kalinowski [b], Fajar Ekaputra [a], Amadeu Anderlin Neto [c], Tayana Conte [c], Dietmar Winkler [a]

[a] Vienna University of Technology
CDL-Flex, Favoritenstr. 9/188
1040 Vienna, Austria
+43 1 58801 18810
<firstname>.<lastname>@tuwien.ac.at

[b] Federal University of Juiz de Fora
NEnC, Rua José Kelmer s/n
36036-330 Juiz de Fora, Brazil
+55 32 2102-3311
kalinowski@ice.ufjf.br

[c] Federal University of Amazonas
IComp, Av. Rodrigo Otávio 6200
69077-000 Manaus, Brazil
+55 92 3305-1193
{neto.amadeu, tayana}@icomp.ufam.edu.br

ABSTRACT

[Context] Experiment planners need to be aware of relevant Threats to Validity (TTVs), so they can devise effective control actions or accept the risk. [Objective] The aim of this paper is to introduce a TTV knowledge base (KB) that supports experiment planners in identifying relevant TTVs in their research context and actions to control these TTVs. [Method] We identified requirements, designed and populated a TTV KB with data extracted during a systematic review: 63 TTVs and 149 control actions from 206 peer-reviewed published software engineering experiments. We conducted an initial proof of concept on the feasibility of using the TTV KB and analyzed its content. [Results] The proof of concept and content analysis provided indications that experiment planners can benefit from an extensible TTV KB for identifying relevant TTVs and control actions in their specific context. [Conclusions] The TTV KB should be further evaluated and evolved in a variety of software engineering contexts.

Categories and Subject Descriptors

D.2.4 [**Software Engineering**]: Software Validation
I.2.4 [**Artificial Intelligence**]: Knowledge Representation

General Terms

Measurement, Experimentation, Theory.

Keywords

Threats to Validity, Controlled Experiment, Systematic Review, Knowledge Engineering, Body of Knowledge.

1. INTRODUCTION

According to Basili *et al.* [2], experimentation in software engineering (SE) supports the advancement of the field through an iterative learning process. In this context, fundamental issues focus on the validity of experiment results [13] and the storage capability of knowledge for learning [3]. Validity is a property of inferences and every experiment faces a range of Threats to Validity (TTVs) [10]. For risk management, experiment planners need to be aware of relevant TTVs in order to properly control or consciously accept those threats [13]. To help researchers in identifying and addressing TTVs, guidelines, generic checklists, and summaries have been presented [10][13]. While these TTV collections provide a good starting point, in general, they do not consider relationships between TTVs and their control actions (e.g., actions to avoid or mitigate risks from TTVs). Therefore, researchers may miss important knowledge in identifying and addressing TTVs while planning experiments.

Empirical Software Engineering (EMSE) experiment reports usually discuss TTVs and control actions, but these reports are isolated in the sense that they do not provide a consolidated discussion on typical TTVs for experiments in the related research area. Therefore, we assume that experiment planners can benefit from a structured overview on TTVs and control actions collected from a sufficiently wide range of actual experimental SE research.

Figure 1 illustrates challenges for stakeholders in the experiment process concerning TTVs: experiment planners use books and reports obtained from digital libraries as support for identifying TTVs and control actions. (1) However, digital libraries and search engines do not support semantic search based on domain concepts. Thus, experiment planners risk missing important TTVs regarding their specific experiment plan. Systematic Literature Review (SLR)/meta researchers can help to extract and analyze relevant knowledge on TTVs and control actions. (2) However, currently most SLRs publish reports but have, in general, no workspace for systematically integrating extracted evidence and sharing it with other researchers to reuse and extend [8].

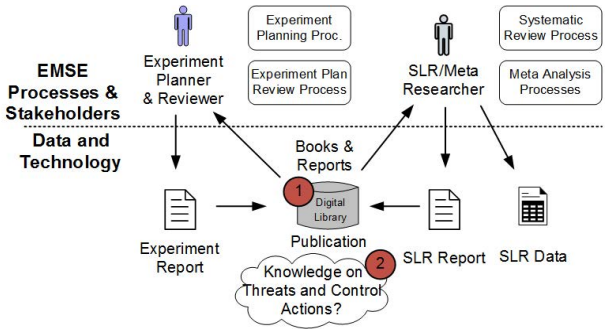

Figure 1. Stakeholder challenges concerning TTVs.

A knowledge base (KB) on TTVs and control actions could facilitate experiment planners with semantic access using domain concepts to explore (a) relevant TTVs related to their research context and (b) information on how those TTVs have been addressed by other researchers. A good TTV KB should be flexible enough to allow adding new evidence and the analysis of data considering SE research areas and domain concepts. These capabilities are difficult to achieve with static SLR reports, because data extracted during SLRs commonly stays in a local archive, not enabling other researchers to easily analyze or extend the extracted data.

ESEM'14, September 18–19, 2014, Torino, Italy.

http://dx.doi.org/10.1145/2652524.2652568

In this paper, we address those challenges by introducing a TTV KB, which extends a KB on controlled experiments [4][5]. Based on requirements coming from EMSE experts, we designed an extensible TTV KB that uses semantic technologies to enable semantic querying for domain concepts (e.g., based on synonyms and related concepts). The TTV KB allows researchers to query for TTVs and control actions reported in specific research areas; for instance, based on experiment domain concepts included in the paper title, TTV description, and control action description.

To populate the TTV KB with a suitable sample of TTVs and control actions, we build on the results of a TTV SLR [1], which identified 206 peer-reviewed papers on controlled SE experiments that report 63 TTVs and 149 control actions. We integrated data on TTVs from the SLR and on TTVs from a widely used EMSE textbook (hereafter referred to as generic TTVs) [13]. An informal proof of concept provided preliminary indications that experiment planners can benefit from an extensible TTV KB for identifying relevant TTVs and control actions in their specific context. An analysis of the TTV KB content supports the need of experiment stakeholders for an overview on TTVs from published research.

The remainder of this paper is organized as follows: Section 2 presents the background. Section 3 motivates the research issues and Section 4 describes how the research issues were addressed. Finally, Section 5 discusses the results and concludes the paper.

2. BACKGROUND

A critical element of empirical research is to analyze and mitigate Threats to Validity (TTVs). Experimenters can draw on a wide range of resources to support experiment planning, in particular, considering TTVs: generic TTV lists in books, e.g., [10][13] and specific TTVs reported in similar experiments.

A recent SLR [1] investigated TTVs and control actions reported in published controlled SE experiments. Initially, this TTV SLR analyzed the experiments selected by Sjøberg *et al.* [12] when investigating the quality of published SE experiments. However, Sjøberg *et al.* [12] focused on papers published between 1993 and 2002. For this reason, the review was extended to gather more data on TTVs, covering also the period 2003 to 2011 [1]. The extension used the same procedure described by Sjøberg *et al.* [12], scanning the same journals and conferences proceedings.

As a result, the TTV SLR [1] identified 206 papers containing experiments and TTVs. Overall, the TTV SLR reported 63 unique TTVs and 149 related control actions. The SLR experience showed that it takes considerable effort to collect and analyze information on TTVs. Table 1 shows an example of an identified specific threat to internal validity (INT-T01) and associated control actions (INT-Cxx). The complete list of identified TTVs and related control actions is available online [1].

Table 1. Threat INT-T01 and its Control Actions [1].

Threat description	Control action description
INT-T01: Differences among subjects related to experience.	• INT-C01: Characterize subjects' experience through a questionnaire. • INT-C03: Assign subjects randomly to groups. • INT-C04: Characterize subjects' experience through a pretest. • INT-C16: Assign treatments randomly to subjects. • INT-C25: Group subjects according to their experience levels.

SLRs have become a widely used and reliable research method [7]. However, SLR reports often provide only discussions on selected research questions, while knowledge collected during the SLR process is usually not publicly available for future extensions and analyses [8]. Thus, making incremental evidence aggregation [10] more complex than necessary. To address this risk the "Systematic Knowledge Engineering" approach [4] discusses how to design the data extraction in the SLR process to enable the import into a KB that allows structured access to collected evidence in an extensible way for members of the scientific community.

3. RESEARCH ISSUES

To efficiently collect and analyze information on TTVs and control actions, we propose a TTV KB to provide semantic access to relevant TTVs. The TTV KB should allow experiment planners getting a quick overview on TTVs in similar experiments, as a sound basis for discussing TTVs and control actions in their context. To cope with new kinds of experiments, the TTV KB should be extensible, enabling researchers in the community to provide and search for TTVs and control actions relevant in their context.

The strategy of building a reusable and extensible TTV KB addresses the challenges depicted in Figure 1. (1) Stakeholders can formulate needs that become semantic queries to the KB; therefore, an important part of the TTV KB is a glossary to provide definitions for key domain concepts, including synonyms and related concepts. (2) TTV data from individual experiments or extracted during SLRs can be systematically integrated into the KB. In addition, data can be extracted from the KB for further analyses. Introducing the TTV KB requires a "Knowledge Engineer", who administers the TTV KB and provides querying and data import/export facilities. From this strategy, we derive the following Research Issues (RIs).

Research Issue RI-1: TTV KB Stakeholder Needs. The starting point for developing a TTV KB is: What are the most relevant needs and requirements for a TTV KB prototype from the viewpoint of EMSE process stakeholders?

Research Issue RI-2: TTV KB Data Model & Content Analysis. Which data elements are necessary to address the most relevant queries from EMSE stakeholders on TTVs and control actions? How can the TTV KB content bring new insight on the use of TTVs in SE experiments compared to the available textbooks?

4. THE TTV KB

This section presents the TTV KB concept, addressing the RIs identified in Section 3.

4.1 TTV KB Stakeholder Needs

Based on informal discussions with SE experiment planners we found that, when considering which TTVs could be relevant for them, they refer to generic TTVs from textbooks and to reports on similar experiments. Therefore, an important question is how to define similarity and how to search for similar experiment reports. An option is to search for experiments in the same research area, which are likely to cover similar kinds of TTVs. For grouping experiment reports by research area we decided to use the 16 chapters of the Software Engineering Body of Knowledge (SWEBOK) [6] (e.g., software quality, software design, or SE management) and also more specific SE topics, hereafter referred to as BOK Topics, derived from the content of the experiment reports (e.g., software inspection, UML, or pair programming).

In a pilot study we identified relevant stakeholder queries by providing an initial set of 10 queries to 10 EMSE experts from six different research groups (located in Austria, Brazil, Spain and Sweden). The selection of the most relevant queries was based on a limited budget of value points, which each stakeholder could spend on identifying query candidates. Then the query candidates were sorted in descending order by the total number of points. The three most relevant Stakeholder Queries (SQ) were:

SQ1: Which research areas have been addressed by the experimental studies reporting TTVs? This query, of special interest to meta-researchers, provides an overview on the experimental research reporting TTVs and on the KB content.

SQ2: Which TTVs have been reported in a given research area (filtered by TTV type and experiment domain concept)? This query allows experiment planners to identify TTVs reported by experiments in similar research areas, optionally filtering by type (internal, external, construct, and conclusion [13]) and experiment domain concepts (e.g., subject selection, hypotheses, instrumentation, and statistical test) and their synonyms.

SQ3: Which control actions have been reported for a specific TTV? This query allows experiment planners to see how other researchers have addressed a selected TTV.

Informal requirements beyond the queries concerned the interfaces to a TTV KB: a query interface (e.g., via a web user interface), and efficient data import/export interface (e.g., via spreadsheets).

4.2 TTV KB Data Model & Content Analysis

We analyzed and extended the data model designed in our previous work [4], where an "EMSE SI research KB" was designed to query for empirical evidence from software inspection experiments collected in a SLR-based process. This data model contained relevant data elements on experiments (based on experimental concepts described in [13]) linked to BOK Topics and was, therefore, a good foundation to design the TTV KB and address the TTV KB requirements. Figure 2 shows part of the TTV KB data model in UML notation. Yellow classes represent newly added entities; light blue classes represent relevant preexisting EMSE SI KB entities. The complete KB data model and remaining experiment-related entities is available online[1].

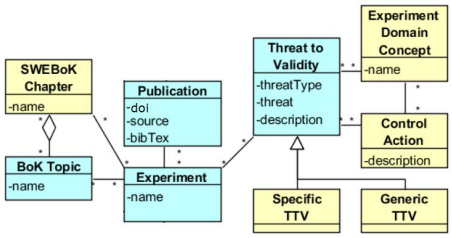

Figure 2. TTV KB data model based on Biffl *et al.* [4].

The knowledge engineer (KE) implemented the TTV KB prototype[1], based on the data model, as an ontology using the Protégé[2] framework. The KE also provided a web prototype for querying the TTV KB and an interface to import data from spreadsheets. The queries were designed using SPARQL[3] ontology query language. To support semantic querying, relevant domain concepts (and their synonyms) were defined in a glossary, also available online[1]. The set of queries is extensible, enabling researchers to reuse the knowledge for applying analyses according to their specific research goals.

To populate the TTV KB, we used the TTV SLR [1] spreadsheet containing extracted data on experiments, related publications, TTVs, and control actions. However, to address the requirements (RI-1) properly by using the KB data model (RI-2) we had to extend the spreadsheet content. The extension comprised: (a) adding information on SWEBOK Chapters and BOK Topics related to each of the 206 experiments; (b) adding information on experiment domain concepts related to each TTV and control action; and (c) adding the generic TTVs reported by Wohlin *et al.* [13] (most-cited TTV identification source, cited as source by 47 experiments) and mappings between the specific and generic TTVs. The mappings were established using a peer-review process. For each specific TTV, a researcher checked if there is a relationship to generic TTVs. Two researchers analyzed all specific TTVs until reaching consensus. If consensus was not reached, a third researcher decided whether the relationship is valid or not.

An informal proof of concept on using the TTV KB (involving two sites located in Austria and Brazil exploring the TTV KB to identify TTVs and control actions in the areas of "software inspection" and "UML"), provided preliminary indications that experiment planners can benefit from an online extensible TTV KB [5]. The queries were considered useful to identify TTVs reported in similar research areas and on specific experiment domain concepts (e.g., subject selection and training).

We analyzed the TTV KB content to check whether a TTV KB can improve experiment planners awareness on relevant TTVs beyond the information available in a standard textbook on experimentation. The TTV KB content analysis showed a high variation in the number of specific TTVs reported in an experiment (see Figure 3). The median was 8 and the maximum was 29 specific TTVs. This high variation seems to indicate a need for additional support with an overview on likely relevant TTVs on top of a structured process to identify TTVs for an experiment design as discussed in textbooks, such as [10][13].

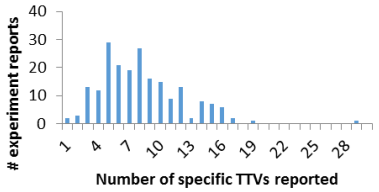

Figure 3: Number of TTVs reported in SE experiments.

There are only a few SE research areas with a significant number of experiments reporting TTVs. The four top-listed SWEBOK chapters (software quality, SE methods, software design, and SE models and methods) are each associated to 35 to 85 experiments. Surprisingly, most SWEBOK chapters are associated to five or less experiment reports. However, even in research areas with a medium-sized set of experiment reports, this set discusses a wide range of specific TTVs, which can be useful to consolidate (e.g., the BOK topic "UML" associated to just 15 experiments still covers 64% of the specific TTVs and 75% of the generic TTVs found in all 206 experiments). The content analysis showed that a small set of TTVs has been reported in many experiments and the relevance of TTVs may vary notably for different research areas.

[1] TTV KB Prototype: http://cdlflex.org/prototypes/ske/threats
[2] Protégé: http://protege.stanford.edu/
[3] SPARQL: http://www.w3.org/TR/rdf-sparql-query/

Finally, the TTV KB content analysis showed that more than half of the 206 experiment reports discussed specific TTVs that could not conservatively be mapped to generic TTVs in the most-cited textbook [13]. Furthermore, five of the generic textbook TTVs were not directly reported in any of the 206 experiment reports. Given this overall scenario, we believe that the TTV KB content analysis supports the notion of a need to provide experiment stakeholders with a consolidated overview on TTVs from published experimental research in the different SE areas.

5. DISCUSSION AND CONCLUSIONS

In this paper we argue that experiment planners can benefit from a structured overview on TTVs and control actions collected from actual experimental SE research, complementing other valuable sources for identifying TTVs, such as guidelines and generic checklists [10][13]. We defined a strategy of providing such overview by introducing a semantic TTV KB populated with data from a TTV SLR [1], extended with information on the research areas and experiment domain concepts. Concerning the TTV KB content quality, the SLR's primary studies were selected based on strict quality criteria and data extensions (e.g., mapping specific to generic TTVs) were added using a peer-reviewed process. The resulting TTV KB prototype is extensible and available online[1].

An informal proof of concept provided preliminary indications that experiment planners can benefit from an online extensible TTV KB for identifying relevant TTVs and control actions in their specific context. The TTV KB content analysis indicates that TTV data from actual experimental research may be useful to complement other TTV identification sources (for instance, more than half of the 206 experiments discussed specific TTVs that could not be mapped to generic TTVs described in a relevant standard EMSE textbook [13]). Thus, we argue that knowledge on TTVs and control actions reported in experimental practice should be made available efficiently to experiment planners.

Based on the TTV KB prototype, experts in the EMSE community collaborating with the KE can incrementally improve the TTV KB content for their research areas and export research data for their advanced analyses. Concerning content extension, it is noteworthy that the TTV KB allows partial data integration. Thus, researchers could extract other experiment-related data from experiment reports (e.g., experiment design, hypotheses, and response variables) and integrate them to increase the KB content, significantly extending the querying capabilities.

As next steps, we plan conducting further evaluations on the TTV KB to strengthen its usefulness and opening it to the community for usage and content extension. These evaluations include a wider survey on needs and requirements and experimental evaluations on the impact of using a TTV KB to support experiment planners. Regarding opening the TTV KB to the community, as recently stated by Basili [3], SE requires a "community supported living experience base". In line with this argument, we believe that a major concern to enable a growing KB is community involvement for frequent updates and quality assurance. Thus, collective intelligence functions [9], enabling to annotate, discuss, evaluate, and correct content are an important foundation for future work.

ACKNOWLEDGMENTS

This work was supported by the Christian Doppler Forschungsgesellschaft, the Federal Ministry of Economy, Family and Youth and the National Foundation for Research, Technology and Development - Austria, and CNPq - Brazil.

REFERENCES

[1] A. Anderlin-Neto and T. Conte, "Threats to Validity and their Control Actions – Results of a Systematic Literature Review", *Technical Report, USES TR-USES-2014-0002,* Federal University of Amazonas, March 2014. Available at http://uses.icomp.ufam.edu.br/attachments/article/42/TR-USES-2014-0002.pdf

[2] V.R. Basili, R. Selby and D. Hutchens, "Experimentation in Software Engineering," In: *IEEE Trans. on SE*, 12(7), 1986.

[3] V.R. Basili, "A personal perspective on the evolution of empirical software engineering" In: *J. Munch and K. Schmid (editors), Perspectives on the Future of Software Engineering*, pp. 255–273. Springer, 2013.

[4] S. Biffl, M. Kalinowski, F. J. Ekaputra, E. Serral, and D. Winkler, "Building Empirical Software Engineering Bodies of Knowledge with Systematic Knowledge Engineering", In: *26th Int. Conf. on Soft. Eng. and Know. Eng. (SEKE)*, 2014.

[5] S. Biffl., M. Kalinowski, F.J. Ekaputra, A. Anderlin- Neto, T. Conte, and D. Winkler, "Towards a Semantic Knowledge Base on Threats to Validity and Control Actions in Controlled experiment", *Technical Report, IFS-CDL 14-04*, Vienna University of Technology, March 2014, Available at http://qse.ifs.tuwien.ac.at/publication /IFS-CDL-14-04.pdf.

[6] P. Bourque and R.E. Fairley (eds.), "SWEBOK v3.0 – Guide to the Software Engineering Body of Knowledge", *IEEE Comp. Soc*., 2014.

[7] S. MacDonell, M. Shepperd, B. Kitchenham, and E. Mendes, "How Reliable Are Systematic Reviews in Empirical Software Engineering?," In: *IEEE Trans. on SE*, 36(5) pp. 676–687, 2010.

[8] E. Mendes, M. Kalinowski, D. Martins, F. Ferrucci, and F. Sarro, "Cross- vs. Within-Company Cost Estimation Studies Revisited: An Extended Systematic Review", *In: Proc. of the 18th EASE Conference*, 2014.

[9] J. Musil, A. Musil, and S. Biffl, "Elements of Software Ecosystem Early-Stage Design for Collective Intelligence Systems," In: *Proc. of Int'l Wsh on Soft. Ecosystem Arch.. Colocated with the 9th Joint Meeting of ESEC/FSE*, 2013.

[10] P.S.M. Santos and G.H. Travassos, "On the Representation and Aggregation of Evidence in Software Engineering: A Theory and Belief-based Perspective," *Electron. Notes Theor. Comput. Sci.*, Vol. 292, pp. 95–118, 2013.

[11] W.R. Shadish, T.D. Cook, and D.T. Campbell, "Experimental and quasi-experimental designs for generalized causal inference," *Houghton Mifflin*, 2002.

[12] D.I.K. Sjøberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N.K. Liborg and A.C. Rekdal, "A survey of controlled experiments in software engineering*", IEEE Trans. on Soft. Eng.*, 31(9), pp. 733-753, 2005.

[13] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, "Experimentation in Software Engineering," *Springer*, 2012.