Evaluating the efficiency of using a search-based automated model merge technique

Conference Paper · October 2018

DOI: 10.1109/VLHCC.2018.8506512

5 authors, including:

Ankica Barisic
NOVA-LINCS
28 PUBLICATIONS **114** CITATIONS
SEE PROFILE

Csaba Debreceni
Budapest University of Technology and Economics
17 PUBLICATIONS **81** CITATIONS
SEE PROFILE

Vasco Amaral
Universidade NOVA de Lisboa
163 PUBLICATIONS **1,004** CITATIONS
SEE PROFILE

Miguel Goulão
Universidade NOVA de Lisboa
101 PUBLICATIONS **581** CITATIONS
SEE PROFILE

Evaluating the efficiency of using a search-based automated model merge technique

Ankica Barišić*, Csaba Debreceni†‡, Daniel Varro†‡§, Vasco Amaral* and Miguel Goulão*
*NOVA LINCS, DI, FCT/UNL, Quinta da Torre 2829-516, Caparica, Portugal
†MTA-BME Lendület Cyber-Physical Systems Research Group, Budapest, Hungary
‡Budapest University of Technology and Economics, Budapest, Hungary
§McGill University, Montreal, Quebec, Canada

***Abstract*—Model-driven engineering relies on effective collaboration between different teams which introduces complex model management challenges. DSE Merge aims to efficiently merge model versions created by various collaborators using search-based exploration of solution candidates that represent conflict-free merged models guided by domain-specific knowledge.**

In this paper, we report how we systematically evaluated the efficiency of the DSE Merge technique from the user point of view using a reactive experimental Software engineering approach. The empirical tests included the involvement of the intended end users (i.e. engineers), namely undergraduate students, which were expected to confirm the impact of design decisions. In particular, we asked users to merge the different versions of the same model using DSE Merge when compared to using Diff Merge. The experiment showed that to use DSE Merge participant required lower cognitive effort, and expressed their preference and satisfaction with it.

***Index Terms*—Domain-Specific Languages, Usability Evaluation, Software Language Engineering**

I. Introduction

Model Driven Engineering (MDE) of critical cyber-physical systems (like in the avionics or automotive domain) is a collaborative effort involving heterogeneous teams which introduces significant challenges for efficient model management. While existing integrated development environments (IDEs) offer practical support for managing traditional software like source code, models as design artefacts in those tools are inherently more complex to manipulate than textual source code.

Industrial collaboration relies on version control systems (like Git or SVN) where differencing and merging artefacts is a frequent task for engineers. However, model difference and model merge turned out to be a difficult challenge due to the graph-like nature of models and the complexity of certain operations (e.g. hierarchy refactoring) that are common today.

In the paper, we focus on an open source tool developed within the MONDO European FP7 project [1] called DSE Merge [2]. DSE Merge presents a novel technique for search-based automated model merge [3] which builds on off-the-shelf tools for model comparison, but uses guided rule-based design space exploration (DSE) [4] for merging models. In general, rule-based DSE aims to search and identify various design candidates to fulfil specific structural and numeric constraints. The exploration starts with an initial model and systematically traverses paths by applying operators. In this context, the results of model comparison will be the initial model, while target design candidates will represent the conflict-free merged model.

978-1-5386-4235-1/18/$31.00 ©2018 IEEE

While existing model merge approaches detect conflicts statically in a preprocessing phase, this DSE technique carries out conflict detection dynamically, during exploration time as conflicting rule activations and constraint violations. Then multiple consistent resolutions of conflicts are presented to the domain experts. This technique allows incorporating domain-specific knowledge into the merge process with additional constraints, goals and operations to provide better solutions.

II. Evaluation approach

Practitioners are still experiencing problems to adopt modelling techniques in practice. Among other factors, developers seem to underestimate the importance of properly aligning the developed modelling tooling to support the techniques with the needs of their end users. We argue that this can only be done by properly assessing the impact of using the technique in a realistic context of use by its target domain users. Investment in the *usability evaluation* is justified by the reduction of development costs and increased revenues enabled by an improved effectiveness and efficiency [5].

Existing Experimental Software Engineering techniques [6] combined with Usability Engineering [7] can be adopted to support such evaluations [8]. This includes the application of experimental approaches, testing empirically with humans, and using systematic techniques to confirm the impact of design decisions on the usability of the developed tools.

Language usability can be defined as the degree to which a language can be used by specific users to meet their needs to achieve particular goals with effectiveness, efficiency and satisfaction in a particular context of use (adapted for the specific case of languages from [9]).

User-centered design (UCD) [10], [11] can contribute to more usable DSLs. For example, [12] presented an innovative visualisation environment, which eases and makes more effective the experimental evaluation process, implemented with the help of UCD. A visual query system was also designed and implemented following the UCD approach [13].

Conducting language usability evaluations is slowly being recognised as an essential step in the Language Engineering lifecycle [14]. An iterative approach allows us to trace usability

requirements and the impact of usability recommendations throughout the DSL development process [8].

III. Experiment

A. Experiment Preparation

The subjects with a different level of modelling expertise were selected to participate in experiment execution based on an online survey held before the experiment. Meanwhile, the development team prepared a demo for DSE Merge tool, the tasks and training material, and finally the virtual machine environment. The materials were evaluated during the pilot session that took place before the experiment execution. The participants of the pilot session were two academics that did not participate in the development of the evaluated tool.

Before starting the experiment, decisions have to be made concerning the *context of the experiment*, the *hypotheses under study*, the set of *independent and dependent variables* that will be used to evaluate the hypotheses, the *selection of subjects participating in the experiment*, the *experiment's design and instrumentation*, and also an *evaluation of the experiment's validity* [8]. The outcome of planning is the experimental evaluation design, which should encompass enough details to be independently replicable.

B. Experiment Objective

Our experiment addresses the following research question:

- *How usable is the proposed technique for performing the model merge operations when compared to the alternative?*

In particular we tested the following hypotheses regarding the use of DSE Merge when compared to the alternative:
Engineers can perform model merge operations ...

- *H1:* more effectively, producing correct results (i.e. merged models are of better quality).
- *H2:* more efficiently (i.e. obtained faster merged models).
- *H3:* more satisfactory (i.e. the modelling activity is perceived as more pleasant)
- *H4:* with less cognitive effort (i.e. lower modelling workload)

C. Experiment Context

The planning of the experiment started by defining explicitly the context of use for technology under evaluation, namely DSE Merge tool. The alternative, i.e. baseline support for model merge problem that is suitable for experimental comparison is identified to be the following:

- Diff Merge [15] shows all the changes to the user where the changes have to be applied manually one by one. Its strength is the user-friendly UI which is very intuitive for the novice users.
- EMF Compare [16] is default comparison and merge tool in the Eclipse environment. In each step, the tool shows only a subset of the changes that the user has to apply into the merge model. Its strength is the capability of handling very complex impacts of changes.

The alternative solutions are meant to support software engineers during the model merge process. The additional benefit claimed for the DSE Merge tool is its power to support domain experts in the same process without requiring from these experts a high level of programming expertise. DSE Merge is claimed to empower incorporation of domain-specific knowledge explicitly into the merge process. However, these two benefits can only be evaluated afterwards. This experiment was scoped to the similar context as alternative supports, to confirm its benefits in the familiar context described as follows:

- *User Profile* - target users for this experiment are expected to be software engineers
- *Technology* - all three tools are running over Eclipse IDE. OS during evaluation was Windows 7 on Desktop computer (Intel(R) Core(TM) i5 650@3.2GHz, 8 GB RAM, 19") or Lenovo Thinkpad T61p laptop (Intel T7700@2.4GHz, 4GB RAM, 15.4").The two languages were tested per subject in the same machine.
- *Social and Physical environment* - the tool is expected to be used in a typical office environment, where the user is working individually by the desk using a laptop or desktop computer. Interaction is performed by use of the mouse, keyboard and the monitor.
- *Domain* - the domain chosen for the experiment was the Wind Turbine case study [17] developed by the industrial partners of the MONDO European Project, as it was previously well-defined and understood by our team.
- *Workflow* - due to the existence of the two different versions representing the same instance model, the user needs to find the best merge solution. The problem is more complex depending on the number of conflicts between the models. We defined the task (T0) as representative to problem reasoning based on domain example.

D. Experiment Flow

The experiment took place at the Budapest University of Technology and Economics [18]. The experimental process started by *Learning Session*, during which the subjects filled the *Background* questionnaire. After this they continue to solve the exercises during *Task Session* which was video recorded. Finally, during *Feedback Session* participants filled final questionnaire rating tools that they have used. Figure 1 depicts the flow of activities during the experiment, explicitly shows documents and treatments that were provided to participants, as well as the instruments that were used to collect the data.

1) Training materials: In the *Learning Session* the participants were allowed to ask questions and were provided with:

- the Wind Turbine Control System meta-model.
- the EMF-models demo video describing the use of Eclipse Modelling IDE and model merge problem.

During the *Tool Session* participants were not allowed to ask any question until the session was finished and were provided with following documents for each evaluated tool:

- the Demo video describing the use of the tool through a presentation of the task T0 that was defined in the experimental workflow context.

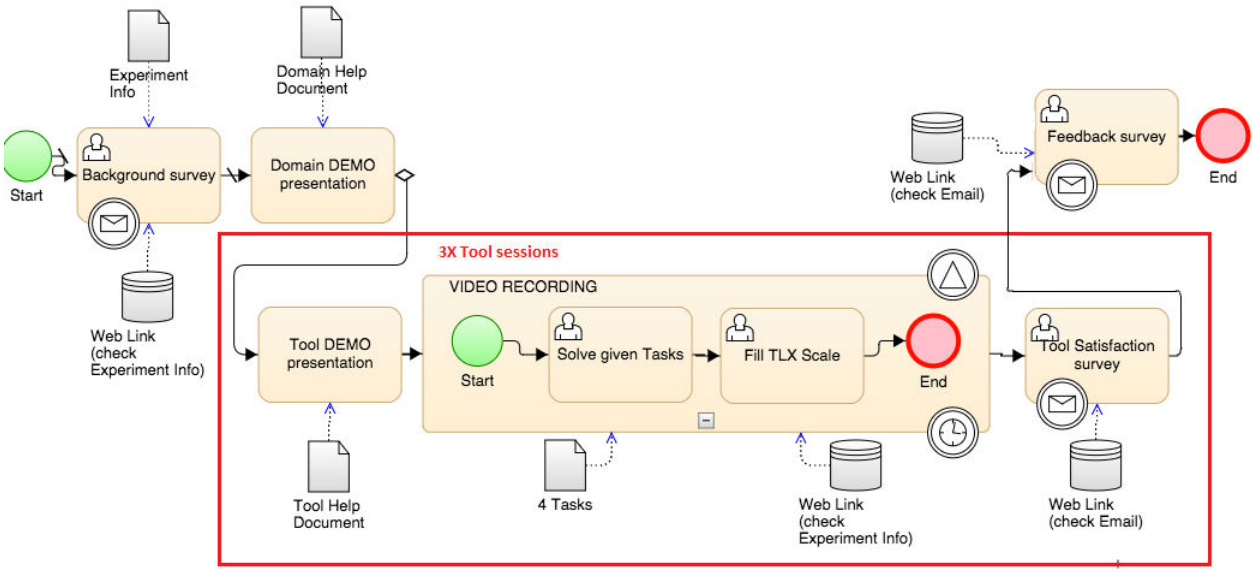

Fig. 1. Experiment treatments

- the Printed document containing explanations and screen-shots presented in the demo video.

During the pilot session, the participants were asked to give the feedback about training directly on the printed materials. Time was estimated to be 10 minutes for *Learning Session*, while 5 minutes for each *Tool Sessions*.

2) Experiment instruments and measurements: These factors are presented in Table I.

TABLE I
INSTRUMENTS AND SCALES

	Instrument	*Value*
Profile	Availability Form, Background Questionnaire	[0-5]
Duration	Video recording	mm:ss
Success	Eclipse project delivery	[0-1]
Cognitive Effort	NASA TLX Scale	[0-1]
Satisfaction	Satisfaction Questionnaire	[(-1)-1]
Preference	Feedback Questionnaire	0 or 1

The data for calculating the *Profile* factor was collected through Availability and Background questionnaire. The *Profile* is influenced by experience in: Modelling; Education and programming; EMF Compare tool; Diff Merge tool; DSE Merge tool; and Wind Turbine metamodel. The *Profile* score (scaling from 0-5) was calculated as the average of all six Experience factors, to which it was added the value of 1 in a case that person had relevant Industry experience. In another case the person was assumed to be Academic.

Duration reflects the actual time taken to solve the tasks and was captured through video analysis.

Success reflects the multiplication of the Success Factor and the Quality Factor. The Quality Factor is defined for each task separately with the following values: 0, 0.25, 0.5, 0.75, 1. These predefined values reflect the number of conflicts that were resolved in contrast with a number of possible correct solutions delivered. The Success Factor took the following values: 1 (if the project reflects the set of correct solutions and is delivered with success); 0.5 (project delivered but is not reflecting the set of correct solutions); 0 (no project delivery). The time to complete the 4 tasks was limited.

Cognitive Effort reflects the participant's workload during solving the task and is measured by a NASA TLX Scale [19].

The *Satisfaction* scale was reflecting average values regarding the following factors: Easy of Use; Confidence; Readability and Understandability of User Interface; Expressiveness; Suitability for complex problems; and Learnability.

The *Preference* factor reflects a clear preference toward one of the tools used based on a subset of Satisfaction criteria, that is annulled if in conflict with the same factor collected using Satisfaction Questionnaire.

All defined instruments were used during the pilot session. In an interview, the evaluator collected the suggestions and doubts regarding the surveys developed for the experiment.

3) Tasks: The representative tasks, of different level of complexity (see Table II), were defined and analysed to be used during experiment execution. During the pilot session, the cognitive effort for each task was estimated to be similar. Time was ranging between 3-5 minutes, while the success rate was high and it was a bit lower for more complex tasks.

TABLE II
TASK VALIDATION

Task	*Model Size*	*Change Size*	*Solutions*	*Cognitive Effort*	*Time*	*Success*
T1	Small	4	2	25.83	3:32	1
T2	Small	12	8	28.61	4:59	1
T3	Big	6	2	20.55	3:18	0.88
T4	Big	54	>million	24.02	4:27	0.83

Based on the obtained results and opinions of the participants during Pilot Session, Diff Merge was found to be a better alternative to DSE Merge for the designated tasks Thus EMF Compare was excluded from the main experiment and left to be optional for the participants after solving the exercises using

TABLE III
COMPARING *Diff Merge* WITH *DSE Merge* - WELCH T TEST

		Diff Merge	DSE Merge	M Diff	S Err Diff	Lower CI	Upper CI	t	df	Sig. (2-tailed)
H1	Success	0.82	0.90	-0.08	0.06	-0.20	0.03	-1.47	22.31	0.16
H2	Duration	1355.71	1289.36	66.36	188.90	-324.39	457.10	0.35	23.02	0.73
H3	**Satisfaction**	0.04	0.27	**-0.23**	0.09	-0.41	-0.05	-2.66	27.00	**0.01**
	- Frustration	58.00	51.43	6.57	9.68	-13.32	26.46	0.68	26.07	0.50
	- EasyToUse	0.00	0.50	**-0.50**	0.19	-0.90	-0.10	-2.58	25.63	**0.02**
	- Confidence	-0.03	0.32	**-0.35**	0.18	-0.73	0.02	-1.95	26.97	**0.06**
	- User Interface	0.07	0.21	-0.15	0.19	-0.54	0.25	-0.77	26.68	0.45
	- Expressiveness	0.20	0.57	**-0.37**	0.14	-0.66	-0.09	-2.68	26.42	**0.01**
	- Suitability	-0.13	0.32	**-0.45**	0.21	-0.88	-0.03	-2.19	26.62	**0.04**
	- Learnability	0.27	0.68	**-0.41**	0.18	-0.78	-0.05	-2.31	27.00	**0.03**
H4	**TLX**	65.31	53.09	**12.22**	5.93	0.03	24.41	2.06	26.03	**0.05**
	- Mental Demand	76.33	67.86	8.48	8.12	-8.46	25.42	1.04	19.97	0.31
	- Physical Demand	28.00	25.36	2.64	11.21	-20.35	25.64	0.24	26.99	0.82
	- Temporal Demand	46.67	51.07	-4.40	10.40	-25.79	16.98	-0.42	26.11	0.68
	- Performance	59.00	57.50	1.50	10.31	-19.68	22.68	0.15	26.30	0.89
	- Effort	66.67	58.21	8.45	7.94	-7.97	24.88	1.07	22.95	0.30

the evaluated tools. The experimental groups were divided into two (G1, G2). G1 received the first *Tool Session* for Diff Merge and then DSE Merge. G2 had the opposite sequence of G1.

IV. RESULTS

Subjects background - Out of 15 participants, 8 of them were from industry and 7 from academia. Most participants were experienced in programming and modelling, but none of them had experience in the Wind Turbine domain. Some participants had previous experience with alternative tools (mostly with EMF Compare), but only one had some basic knowledge of DSE Merge.

TABLE IV
SUBJECT BACKGROUND

	Total	*G1*	*G2*
Number of participants	15	6	9
Profile	1.65	1.92	1.39
Industry	56%	67%	44%

Comparative results - We compare the results for DSE Merge and Diff Merge in Table III. For each measured attribute, we present its mean value with Diff Merge, its mean value with DSE Merge, the mean difference between both, the standard error of that difference, the 95% confidence interval lower and upper boundaries, the Welch t-test statistic, its degrees of freedom and $p-value$. For hypothesis $H1$, although on average there was a slight improvement, we found no statistically significant difference between using both languages and, therefore, no evidence supporting the hypothesis that developers would achieve a higher success with DSE Merge. For hypothesis $H2$, although on average participants were slightly faster with DSE Merge, we found no statistically significant evidence supporting the hypothesis that the task would be performed more efficiently with DSE Merge when compared to Diff Merge. For $H3$, there was a statistically significant difference supporting the hypothesis that using DSE Merge leads to a higher satisfaction than using Diff Merge. This improvement was statistically significant concerning ease of use, confidence, expressiveness, suitability and learnability, with no significant difference concerning frustration or user interface. Finally, concerning $H4$, overall, we found evidence supporting the hypothesis that the overall cognitive effort (NASA TLX global score) using DSE Merge was lower than using Diff merge. The difference is not attributable to any of the individual TLX scores. Finally, from the feedback questionnaire, we obtained the Preference factor of 11 for DSE Merge, while Diff Merge was only rated 1.

Threats to validity - Concerning the selection of the participants, they were all recruited in the same university. This creates a selection validity threat, as they may not fully represent the target population of DSE Merge. Besides, the sample size is relatively small. Replications of this evaluation should be independently conducted at other sites to mitigate these threats. Two other potential threats were hypothesis guessing where participants try to guess the hypotheses under study and change their behaviour as a result of it, and the experimenter's expectations. However, both the experimental evaluation and subsequent data analysis were conducted by researchers external to the development team of DSE Merge, thus mitigating both threats.

V. CONCLUSION

The results of the presented empirical study show that DSE Merge has clear advantages regarding the satisfaction (H3) of their users and the cognitive effort (H4) required to use it.

As future work, we plan to extend this study to subject modellers from the community of both practitioners and academics from outside the Budapest University. For that, we will make use of crowdsourcing platforms. This will allow us to improve both the statistical relevance of this study as well as to minimise the previously identified threat validity of the subjects representativity.

ACKNOWLEDGMENTS

The authors thank COST Action IC1404 Multi-Paradigm Modeling for Cyber-Physical Systems (MPM4CPS) H2020 Framework, as well as NOVA LINCS Research Laboratory (Grant: FCT/MCTES PEst UID/ CEC/04516/2013) and Project DSML4MA (Grant: FCT/MCTES TUBITAK/0008/2014).

REFERENCES

[1] MONDO, "Scalable modelling and model management on the cloud," project. [Online]. Available: www.mondo-project.org/,accessed: 07.11.2015

[2] C. Debreceni, I. Ráth, D. Varró, X. D. Carlos, X. Mendialdua, and S. Trujillo, "Automated model merge by design space exploration," in *Fundamental Approaches to Software Engineering - 19th International Conference, FASE 2016*, ser. LNCS, vol. 9633. Springer, 2016, pp. 104–121.

[3] M. Kessentini, W. Werda, P. Langer, and M. Wimmer, "Search-based model merging," in *Proceedings of the 15th annual conference on Genetic and evolutionary computation*. ACM, 2013, pp. 1453–1460.

[4] A. Hegedus, A. Horváth, I. Ráth, and D. Varró, "A model-driven framework for guided design space exploration," in *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*. IEEE Computer Society, 2011, pp. 173–182.

[5] A. Marcus, "The ROI of usability," in *Cost-Justifying Usability*, Bias and Mayhew, Eds. North- Holland: Elsevier, 2004.

[6] V. R. Basili, "The role of controlled experiments in software engineering research," in *Empirical Software Engineering Issues. Critical Assessment and Future Directions*, ser. LNCS, V. R. Basili, D. Rombach, K. Schneider, B. Kitchenham, D. Pfahl, and R. Selby, Eds. Springer Berlin / Heidelberg, 2007, pp. 33–37.

[7] J. Nielson, *Usability Engineering*. AP Professional, 1993.

[8] A. Barišić, V. Amaral, and M. Goulão, "Usability Driven DSL development with USE-ME," *Computer Languages, Systems and Structures (ComLan)*, vol. ISBN 1477-, 2017.

[9] International Standard Organization, "ISO/IEC FDIS 25010:2011 systems and software engineering – systems and software quality requirements and evaluation (SQuaRE) – system and software quality models," March 2011. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=35733

[10] D. A. Norman and S. W. Draper, "User centered system design," *Hillsdale, NJ*, 1986.

[11] K. Vredenburg, J.-Y. Mao, P. W. Smith, and T. Carey, "A survey of user-centered design practice," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2002, pp. 471–478.

[12] M. Angelini, N. Ferro, G. Santucci, and G. Silvello, "VIRTUE: A visual tool for information retrieval performance evaluation and failure analysis," *Journal of Visual Languages & Computing*, vol. 25, no. 4, pp. 394–413, 2014.

[13] E. Bauleo, S. Carnevale, T. Catarci, S. Kimani, M. Leva, and M. Mecella, "Design, realization and user evaluation of the SmartVortex Visual Query System for accessing data streams in industrial engineering applications," *Journal of Visual Languages & Computing*, vol. 25, no. 5, pp. 577–601, 2014.

[14] T. Kosar, M. Mernik, and J. Carver, "Program comprehension of domain-specific and general-purpose languages: comparison using a family of experiments," *Empirical Software Engineering*, vol. 17, no. 3, pp. 276–304, 2012.

[15] "EMF Diff/Merge," https://wiki.eclipse.org/EMF_DiffMerge, accessed: 2018-07-25.

[16] "EMF compare," https://www.eclipse.org/emf/compare/, accessed: 2018-07-25.

[17] A. Gómez, X. Mendialdua, G. Bergmann, J. Cabot, C. Debreceni, A. Garmendia, D. S. Kolovos, J. de Lara, and S. Trujillo, "On the opportunities of scalable modeling technologies: An experience report on wind turbines control applications development," in *Modelling Foundations and Applications - 13th European Conference, ECMFA 2017,*, ser. LNCS, vol. 10376. Springer, 2017, pp. 300–315.

[18] A. Barišić, "STSM Report: Evaluating the efficiency in use of search-based automated model merge technique," in *Multi-Paradigm Modelling for Cyber-Physical Systems (MPM4CPS)*, no. COST action IC1404. European cooperation in science and technology, 2016. [Online]. Available: http://mpm4cps.eu/STSM/reports/material/STSM_report-Ankica_Barisic.pdf

[19] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Advances in psychology*, vol. 52, pp. 139–183, 1988.