An Empirical Study Concerning Ontology-Driven Modularization and Ontology-Neutral Modularization Techniques

Michaël Verdonck[1], Maria Hedblom[2], Giancarlo Guizzardi[2], Guylerme Figueiredo[3], Frederik Gailly[1] and Geert Poels[1]

[1]Faculty of Economics and Business Administration, Ghent University, [2] Free University of Bozen-Bolzano; [3]Federal University of Espírito Santo, Brazil

{michael.verdonck, frederik.gailly, geert.poels}@ugent.be; {maria.hedblom, giancarlo.guizzardi}@unibz.it; guylerme@gmail.com

Abstract. In this paper, we describe the hypothesis and the design of an empirical study to investigate the impact of adopting an ontology-driven modularization (ODM) technique and an ontology-neutral modularization (ONM) technique on the comprehension of the resulting modules of the conceptual model. More specifically, we intend to measure the impact of these modularization techniques on the comprehension and understandability of the modules by novice modelers – a category of users that greatly benefit from such modularization techniques. While much effort in ontology-driven modularization has been devoted into maintaining the semantic aspects of the original conceptual model during the modularization process, this research effort aims to contribute by performing an empirical validation of such techniques. Since little empirical research has yet been conducted in this area, we plan to perform an experimental study in order to capture the differences between adopting an ontology-driven modularization technique and an ontology-neutral modularization technique.

1 Introduction

Conceptual modeling, in all its various forms, plays an important role in representing and supporting complex human design activities. Conceptual modeling was introduced to increase understanding and communication of a system or domain among stakeholders and can be described as the activity of representing aspects of the physical and social world for the purpose of communication, learning and problem solving among human users [1]. However, as the complexity of these domains increases, so does the size and complexity of the models that represent them. Such increased complexity is problematic in sensitive domains such as finance and healthcare, where conceptual models play a fundamental role in different types of critical semantic interoperability tasks. Therefore, it is essential that domain experts are able to understand and accurately reason with the content of these models. Moreover, the human capacity for processing unknown information is very limited, containing bottlenecks in visual short-term memory and causing problems to identify and hold stimuli [2].

In order to cope with these challenges, certain complexity management techniques were adopted. One such technique is *Conceptual Model Modularization* (CMM). Broadly speaking, CMM is the process of separating a model into varying degrees of interdependence and independence across, with the primary use to hide the complexity of each part behind an abstraction, often with the benefit of flexibility and variety in use [3]. Or in other words, we have the original 'complex' conceptual model that is being separated into different sub-models – henceforth called *modules* – according to certain extraction criteria. While different CMM techniques exists, a great many of these techniques rely on ontologically-neutral modeling languages [4] such as the Unified Modeling language (UML), Extended-Entity Relationship (EER) diagrams or the Web Ontology Language (OWL). Due to the lack of real-world semantics of these modeling languages, the CMM techniques rely on module extraction criteria that leverage almost entirely on the syntactical properties of the models, i.e. topological ones [5]. As a result, CMM techniques often struggle with finding the adequate criteria for performing the module extraction that divides a conceptual model into distinct modules.

As an alternative approach, ontology-driven modularization techniques were developed. More specifically, these techniques rely on the ontological semantics behind an ontology and leverage these semantics in order to modularize a conceptual model into different modules. Ontologies can be described as a foundational theory, which articulates and formalizes the conceptual modeling grammars needed to describe the structure and behavior of the modeled domain [6]. Adopting an ontology-driven modularization approach has the advantage that it makes systematic use of the real-world (ontological) semantics to propose a structure of modules which preserve all the informational content of the original model. Moreover, the ontology-driven approach can also adopt these ontological semantics as criteria to separate the different modules – something which is not possible with ontology-neutral CMM techniques since they only rely on syntactic criteria. As a result, the inclusion of the ontological semantics of the domain should lead to more comprehensible modules compared to that of ontology-neutral CMM techniques. However, while several research efforts [7, 8] have adopted ontological theories to perform modularization or develop techniques that aim to enhance the modularization process of separating the modules from a conceptual models, there exists little research that actually demonstrates that adopting ontology-driven CMM techniques result in more comprehensible modularized conceptual models.

Therefore, it is the goal of this paper to compare an ontology-driven modularization (ODM) technique to an ontology-neutral modularization (ONM) technique and investigate their effect on the comprehension of the resulting modularized models. To properly measure these effects, we conduct an empirical study. As the foundation for the further development of this paper, we formulate our **research question** as follows: *Are there meaningful differences in the comprehension of the modularized conceptual models between adopting an ontology-driven modularization technique and an ontology-neutral modularization technique?* In order to formulate a proper answer, this broad research question is translated into a hypothesis. The testing hypothesis will be formulated in section 2. Next, we will draft our experimental design that we would apply to test these hypotheses in section 3. Finally, we conclude this paper in section 4.

2 Hypothesis Development

A substantial amount of research has been performed to enhance conceptual model modularization, which has resulted in a series of different approaches and techniques. For instance, several approaches focused on applying syntactic analysis techniques for conceptual model modularization [5, 9]. While such approaches are successful in modularizing a conceptual model, they risk however semantically insensitive extraction results as a consequence of neglecting the meanings of the represented elements in the domain. For instance, the lack of semantic expressiveness results in the inability to differentiate between different sort of types, making it more complex to identify domain relevance for the represented concepts. The research of [8] illustrates this point: take for example a class diagram in which we have an ADDRESS class and several other classes such as EMPLOYEE, ORGANIZATION and CLIENT that are connected to it – since conceptually all these types of entities can have one or more addresses. When this model is converted into for instance a graph (adopted by many different modularization techniques), a purely syntactical technique could consider that the most relevant node in that graph should be ADDRESS. However, concluding that this represents the most relevant concept in the domain would be a mistake. Hence, to overcome these issues, different approaches started to incorporate ontologies to perform conceptual model modularization. The advantages of such ODM techniques is that they incorporate the ontological semantics behind the elements of a conceptual model, which can then be used as criteria to extract the modules from the conceptual model and to preserve the informational content of the original conceptual model [8, 10].

Although the adoption and development of these new ontology-driven techniques have demonstrated to be successful in maintaining the semantics expressiveness of the constructs of a conceptual model during the process of modularization, they often lack an empirical evaluation. Moreover, studies that have performed empirical research concerning the application of ODM techniques have focused more on demonstrating that the adoption of such techniques will lead for instance to enhanced module extraction, better search queries or semantically richer modules [7]. It is then often implied that since ODM techniques provides a higher degree of semantic expressiveness of the extracted modules, they will consequently lead to a better comprehension and understanding by the users of these modules. However – as to the knowledge of the authors – no explicit study exists that empirically validates that adopting an ODM technique will lead to more comprehensible modules – and hence a better understanding of the domain that is being modeled – compared to adopting an ONM technique. Since both types of techniques have a different emphasis (i.e. syntactic vs semantic), they consequently influence the structure and components of the conceptual modules that are extracted from the original conceptual model – which as a result can lead to significant differences in the comprehension of these modules and the domain they are meant to represent.

Therefore, we will empirically investigate the impact of adopting an ODM technique compared to a ONM technique on the comprehension of the resulting conceptual modules. Based upon the assumptions given above, we formulate our **hypothesis** as follows: *The comprehension of conceptual models derived from an ontology-driven modularization technique will be higher compared to the comprehension of conceptual models that are derived from an ontology-neutral modularization technique – given a sufficiently complex modeling task and the subjects being novice modelers.* As formulated in the hypothesis, we will test the comprehension of the conceptual modules on novice modelers, that will have no prior experience of either of the techniques – enabling us to measure the full influence of the modularization approach that is being adopted. Novice modelers are a category of users that greatly benefit from modularization techniques since they reduce the complexity of a conceptual model – therefore making sense to adopt novice modelers as our subjects. More specifically, in our experimental setting, we will create different conceptual modules with both an ODM and an ONM technique, and then present these modules to two distinct group of subjects, who will have to interpret these modules. We will then assess the impact of each technique by evaluating the comprehension between the modules that have been constructed by the respective techniques. Since the principal goal of modularization is to reduce the complexity of complex conceptual models, we will adopt a sufficiently complex modeling domain to effectively perform this comparison.

3 Experimental Design

In this section we will outline our experimental design in order to test the hypotheses above. The experimental design is based upon the work of Wohlin et al. [11]. We first define our variables that will be tested. Next, we specify the selection of our subjects. Finally, we explain the choice of our experimental design type, which is composed of an experiment and a protocol analysis.

3.1 Variable development

Independent Variable: In our study, the independent or affecting variable constitutes of the two different modularization techniques that are going to be adopted to construct the conceptual modules that represent a certain domain. In other words, in our experimental setting we can control if we either assign our test subjects with modules created by an ODM technique or with an ONM technique. More specifically, for the ODM technique, we adopt the OntoUML based modularization technique of Figueiredo et al. [8]. OntoUML is an ontology-driven conceptual modeling language, whose modeling constructs and metamodel constraints reflect the ontological distinctions and axiomatization put forth by the Unified Foundational Ontology [12]. Their proposed modularization technique leverages the ontological semantics behind the modeling constructs of the OntoUML language in order to perform module extraction and to maintain the semantics of the modeling constructs during that abstraction. As for the ONM technique, we adopt the technique of Egyed [9], who developed a modularization technique that performs automatic UML class model abstraction. Based on the combination of several existing techniques, an automatic abstraction algorithm was created that is based on a set of rules that transforms the model into a graph and tries to infer abstractions of the model by means of the connectivity of the nodes. We would like to emphasize that the modules of both techniques are entirely represented through the UML modeling language, making them suitable for comparison.

Dependent variables: Model (or module) comprehension can be measured with different approaches. A distinction is made between efficiency and effectiveness [13]. While effectiveness of a modeling technique is defined by how well it achieves its objective – in our case model comprehension – efficiency is defined by the effort required to adopt the modeling technique. The former can be measured by output measures evaluating the quantity and/or quality of the results; the latter can be measured by a variety of input measures such as time, cost or effort. In our paper, the effectiveness will thus directly measure model comprehension, while the efficiency will measure the cost of effort to comprehend the models.

More specifically, we measure the **effectiveness** of the ontology-driven models with comprehension and problem-solving questions. These output measures are similar to previous research studies [14, 15], where they also compared the comprehension and understandability of different kinds of models that were constructed with different development techniques. While the *comprehension questions* assess a basic level of model comprehension, the *problem-solving questions* are more challenging and target a deeper level of model comprehension from the subjects. The **efficiency** of the ontology-driven models will be measured by: (1) assessing the *amount of time* needed to understand the models, and (2) the amount of effort a subject had to spend to fulfill the tasks related to the ontology-driven models, here expressed as the *ease of interpretation* (EOI). The EOI questions are based on the perceived ease of understanding as applied in the research effort of [16]. The EOI questions are divided in such a way that they measure different aspects of perceived effort during the experiment. More specifically, they assess: (i) the effort in comprehending a specific module; (ii) the effort spent to complete the comprehension questions or the problem-solving questions; and (iii) which aspects of the assignment required the most effort to solve.

Control Variable: Since we will be testing users' comprehension of UML class diagrams, we need to ascertain that all subjects have an equal understanding of the UML modeling language. Therefore, we need to assure that the interpretation of a certain module can be linked to the technique that was applied to construct the module, and not to a limitation of the subject's knowledge of the UML modeling language. As such, we apply a control variable to test every subject's knowledge and understanding of UML, before the start of the experiment.

3.2 Subject Selection

The subjects in our study all were novice conceptual modelers that had some prior experience and education in the domain of conceptual modeling, and that were completing their Masters at Ghent University. We decided to select students as our test subjects since they have no prior knowledge of modularization techniques and can thus be seen as a 'tabula rasa'. Consequently, we could measure the full impact of the modularization technique that is being adopted to create the modules – and to measure the influence these techniques on the comprehension of the modules. Furthermore, all subjects were around the same age (i.e. mid-twenties) with the majority of the subjects having a business-oriented background. This specific selection thus leads to a controlled sample of subjects with the same level of experience in conceptual modeling and with no prior knowledge about any of the modularization techniques that were applied in the empirical study.

3.3 Experimental Design Type

Similar to the research of [17, 18], two empirical studies were designed to test our formulated hypotheses. Our design can be divided into two phases: in the first phase we will conduct a protocol analysis to examine the instruments of our experiment (e.g. comprehension questions, problem-solving questions etc.) with the purpose to examine how they are perceived by our test subjects and that serves as a feedback opportunity to further enhance certain aspects of the experiment. In the second phase, we will perform the experiment itself, which will generate a significant amount of data to test our hypothesis. During this phase we will also conduct an in-depth analysis in order to provide additional insights into the nature of our results. While the experiment will be performed on a larger scale with more subjects, the protocol analysis will be performed with a smaller set of subjects, since the goal of the protocol analysis is not to produce data but to acquire insights and feedback on how subjects perceive the instruments of the experiment.

Protocol Analysis

A protocol analysis is a research method that elicits verbal reports from research participants. The data obtained from a protocol analysis method reveals the mental processes taking place as individuals work on the interpretation of the models. Subjects are required to verbalize their thought processes and strategies, as well to verbalize their answers to the comprehension, problem solving and EOI questions. These verbal reports and the progress of the subjects are closely monitored by the researcher guiding the treatment. A protocol analysis thus allows us to closely monitor the interpretation of the models by the subjects, and carefully register the model comprehension of each individual subject. Therefore, before the experiment has been conducted, we will perform the protocol analysis on a set of test subjects, in the exact way as our experiment will be performed but with the purpose to understand the perception and interpretation of the design of our experiment. Naturally, the experiment will be performed on a new set of subjects, since performing the experiment with the same subjects as for the protocol analysis would alter the results since these subjects have already performed the experiment before.

Experiment

An experiment consists of a series of tests of different treatments. To get the desired results to answer our research question, the series of tests must be carefully planned and designed. The design of our experiment is based upon the testing hypotheses we developed earlier. From these hypotheses, we can derive two treatments: an ODM treatment and an ONM treatment. The series of test in each treatment constitute of the different conceptual modules that are generated by these techniques. Our subjects are thus divided into two different treatments, where each treatment submits the subjects to similar tests where the comprehension of the modules is measured. We have assigned the subjects to these treatments according to the *balancing design principle* [11]. By balancing the treatments, we assign an equal number of subjects to each separate treatment, to arrive at a balanced design. Balancing is desirable since it both simplifies and strengthens the statistical analysis of the data.

The design type of our experiment is a *one factor with two treatments* design, meaning that we compare the two treatments against each other with one independent variable (i.e. model comprehension) – also considering the control variable (understanding of the UML modeling language). Our design will also be completely randomized, meaning that subjects will be allocated randomly to either one of the treatments. Each subject also takes part in only one treatment. Most commonly, the means of the dependent variable for each treatment are compared. We will thus assign scores to the different measures of the dependent variable, i.e. the comprehension questions, the problem-solving questions, the amount of time required to solve the task and the ease of interpretation questions, in order to compare our two different treatments objectively.

4 Conclusion and future research

This paper carefully structures and describes the hypothesis and design of an empirical study that will investigate the impact of adopting an ontology-driven modularization technique and an ontology-neutral modularization technique on the comprehension of the resulting conceptual modules. More specifically, this study intends to measure the impact of these two techniques on the comprehension and understandability of the conceptual modules by novice modelers since they are a common category of users for such techniques. As for the next phase in this research project, we are preparing and identifying subjects to participate in our empirical study.

References

1. Mylopoulos, J.: Conceptual modeling and telos. In: Loucopoulos, P. and Zicari, R. (eds.) Conceptual Modelling, Databases and CASE: An Integrated View of Information Systems Development. Wiley (1992).
2. Moody, D.: The physics of notations: Toward a scientific basis for constructing visual notations in software engineering. IEEE Trans. Softw. Eng. 35, 756–779 (2009).
3. Baldwin, C.Y., Clark, K.B.: Design rules: The power of modularity. MIT press (2000).
4. Guizzardi, G., Baião, F., Lopes, M., Falbo, R.A.: The role of Foundational ontologies for Domain ontology Engineering: an Industrial Case Study in the Domain of oil and Gas Exploration and Production. Int. J. Inf. Syst. Model. Des. 1, 1–22 (2010).
5. Nino, A.V. (Universitat P. de C., Ramon, A.O. ((Universitat P. de C., Samso, M.-R.S. (Universitat P. de C.: A filtering engine for large conceptual schemas. (2013).
6. Wand, Weber, R.: On the ontological expressiveness of information systems analysis and design grammars. Inf. Syst. J. 3, 217–237 (1993).
7. Lozano, J., Carbonera, J., Abel, M., Pimenta, M.: Ontology View Extraction: An Approach Based on Ontological Meta-properties. Proc. - Int. Conf. Tools with Artif. Intell. ICTAI. 2014–Decem, 122–129 (2014).
8. Figueiredo, G., Duchardt, A., Hedblom, M.M., Guizzardi, G.: Breaking into pieces: An ontological approach to conceptual model complexity management. In: Proceedings - International Conference on Research Challenges in Information Science. pp. 1–10. IEEE (2018).
9. Egyed, A.: Automated abstraction of class diagrams. ACM Trans. Softw. Eng. Methodol. 11, 449–491 (2002).
10. Aparicio, J.M.L.: Ontology View : a new sub-ontology extraction method. (2015).
11. Wohlin, C., Runeson, P., Host, M., Ohlsson, M.C., Regnell, B., Wesslen, A.: Experimentation in software engineering. (2012).
12. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. CTIT, Centre for Telematics and Information Technology (2005).
13. Moody, D.L.: The Method Evaluation Model : A Theoretical Model for Validating Information Systems Design Methods. Inf. Syst. J. 1327–1336 (2003).
14. Gemino, A., Wand, Y.: Complexity and clarity in conceptual modeling: Comparison of mandatory and optional properties. Data Knowl. Eng. 55, 301–326 (2005).
15. Burkhardt, J.M., Détienne, F., Wiedenbeck, S.: Object-oriented program comprehension: Effect

of expertise, task and phase. Empir. Softw. Eng. 7, 115–156 (2002).
16. Maes, A., Poels, G.: Evaluating quality of conceptual modelling scripts based on user perceptions. Data Knowl. Eng. 63, 701–724 (2007).
17. Bera, P.: Analyzing the Cognitive Difficulties for Developing and Using UML Class Diagrams for Domain Understanding. J. Database Manag. 23, 1–29 (2012).
18. Burton-Jones, A., Meso, P.P.N.: Conceptualizing systems for understanding: An empirical test of decomposition principles in object-oriented analysis. Inf. Syst. Res. 17, 38–60 (2006).