MANAGING PROVENANCE OF SIMULATION STUDIES

Pia Wilsdorf
Andreas Ruscheinski
Marcus Dombrowsky
Adelinde M. Uhrmacher

Institute of Computer Science
University of Rostock
Albert-Einstein-Straße 22
18059 Rostock, GERMANY

ABSTRACT

Various workflow tools for simulation studies exist that also support reuse and reproducibility, e.g., by tracking provenance information and representing it using a provenance graph. However, depending on the level of granularity, the resulting graph may become enormous in size which makes it difficult for users to draw conclusions directly from the tracked information. Therefore, we employ a variety of aggregation techniques to manage the provenance information based on user requirements. Using a typical simulation case study we discuss applicability and effect of the different reduction techniques.

1 MOTIVATION

Performing simulation studies to gain new insights into a system's behavior has become a common technique in many sciences. However, simulation studies are intricate processes that involve successive model refinements and diverse experiment executions that can be described via life cycles, where each simulation study has to pass through different phases. Various workflow tools exist to support users in the different phases of a simulation study, in particular the execution of individual simulation experiments (Görlach et al. 2011). However, even for entire simulation studies whose main product is the simulation model rather than the simulation data, first workflow approaches have been proposed (Rybacki et al. 2014). Declarative artifact-centric approaches are of particular interest as they make the diverse artifacts of a simulation study (e.g., conceptual model, simulation model, simulation experiment, or simulation data) and the stages they move through as well as the constraints based on which stages are enabled or disabled explicit. In addition to supporting the conduction of simulation studies, many workflow systems also support reproducibility and reusability of the designed models and experiments by documenting the entire process, i.e., by tracking their provenance information on-the-fly. For example, our approach tracks provenance by relating the artifacts' stages to the elements of a provenance graph which allows us to answer detailed user queries on demand. However, using this fine-grained tracking mechanism the resulting provenance graph becomes enormous in size even for relatively small simulation studies. Consequently, appropriate reduction techniques are needed to manage this immense amount of information.

2 MANAGING PROVENANCE INFORMATION

Appropriate provenance management techniques are essential to provide users with the ability to retrieve information from the provenance graph in a clear and helpful way (Allen et al. 2010). Since finding the right level of granularity is difficult, we follow the principle: first to collect information exhaustively, to later filter and aggregate on demand by using query filters, or black boxing techniques (Allen et al. 2010;

Missier et al. 2013). For this purpose we extend our artifact-centric simulation workflow framework with a provenance management component which automatically identifies and transforms graph patterns. There the main challenge is to account for the different types of artifacts, dependencies, and roles.

3 CASE STUDY

For a first proof of concept we conduct a sample simulation study with the objective to build a valid model of a goldfish tank which shall be used to predict the behavior of the goldfish depending on environmental conditions. Therefore, based on a conceptual model (CM) first a model of the fish tank is build and calibrated (M1 calib.) followed by a model of the goldfish behavior (M2 calib.). Next, these two submodules are composed to a model M3 and extended to a model M3'. Finally, the composed simulation model is validated successfully using different validation experiments (E6-8) where the results of three wetlab experiments (W5-7) are reproduced. The generated provenance graph for this simulation study contains over 50 process nodes and more than 80 artifact nodes. Applying appropriate aggregations at different parts of the provenance graph and combining the techniques in a suitable way can significantly reduce the size and highlight the principle idea of the simulation study. Figure 1 depicts the combined use of three black boxing techniques.

Version box: Throughout the conduction of the simulation study users may save versions of their work, e.g., to mark achievements like the calibration or validation of a model. The version box technique can be applied if multiple versions of the simulation study exist and only recent provenance information with regard to the latest "save" are of interest, i.e., processes and artifacts belonging to previous versions can be collapsed into boxes (here V.1 and V.2).

Sequence box: Whenever there are repeated actions on the same artifact, they can be combined into a sequence box, e.g., successive updates of the simulation model, or even simple simulation experiments used for analysis. However, some processes such as the execution of calibration and validation experiments must not be included in a sequence box since they play a special role during the simulation study.

Aggregation box: An aggregation box abstracts a set of artifacts by a single node. It can be applied whenever multiple artifacts of the same type are used or produced by the same process node, or even by a series of processes of the same type, e.g., experiment descriptions in a series of validation experiments.

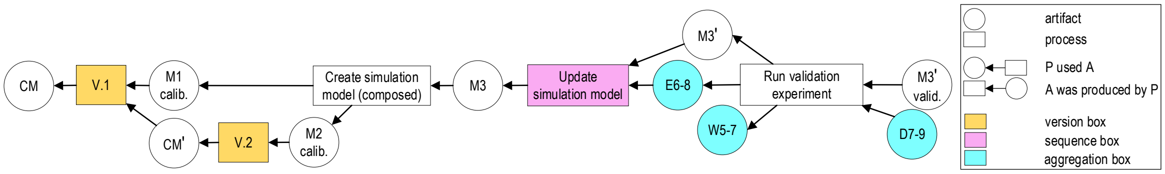

Figure 1: Consecutive application of three different aggregation techniques for a typical simulation study.

ACKNOWLEDGMENTS

This work was supported by the DFG (German Research Foundation) UH 66/18 "GrEASE".

REFERENCES

Allen, M. D., L. Seligman, B. Blaustein, and A. Chapman. 2010. "Provenance Capture and Use: A Practical Guide". *The MITRE Corporation*.

Görlach, K., M. Sonntag, D. Karastoyanova, F. Leymann, and M. Reiter. 2011. *Conventional Workflow Technology for Scientific Simulation*, 323–352. London: Springer London.

Missier, P., J. Bryans, C. Gamble, V. Curcin, and R. Danger. 2013. *Provenance Graph Abstraction by Node Grouping*. Computing Science, Newcastle University.

Rybacki, S., F. Haack, K. Wolf, and A. Uhrmacher. 2014. "Developing Simulation Models: From Conceptual to Executable Model and Back-an Artifact-based Workflow Approach". In *SimuTools*, 21–30.