Towards a PROV Ontology for Simulation Models[*]

Andreas Ruscheinski, Dragana Gjorgevikj, Marcus Dombrowsky, Kai Budde,
and Adelinde M. Uhrmacher

Institute of Computer Science, University of Rostock
Albert-Einstein-Str. 22, 18059 Rostock, GERMANY
andreas.ruscheinski@uni-rostock.de

Abstract. Simulation models and data are the primary products of simulation studies. Although the provenance of simulation data and the support of single simulation experiments have received a lot of attention, this is not the case for simulation models. The question of how a simulation model has been generated requires to integrate diverse simulation experiments and entities at different levels of abstractions within and across entire simulation studies. Based on a concrete simulation model, we will use the PROV Data Model (PROV-DM) and illuminate the benefits of the PROV-DM approach to identify and relate entities and activities that contributed to the generation of a simulation model, thereby taking first steps in defining a PROV-DM ontology for simulation models.

Keywords: Simulation model · Provenance · Simulation study

1 Introduction

Provenance provides "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability, or trustworthiness" [2]. Applying provenance to outcomes of modeling and simulation studies, such as output data and the simulation model, requires to identify central activities and products and to put those into relation. Existing standards like SBML [3] or the ODD protocol [1] document *what has been developed* rather than *how it has been developed.* The provenance of simulation data and the execution of individual simulation experiments, be this single runs, parameter scans, or simulation-based optimization, have been the subjects of major research efforts. Accordingly, different approaches like scripts, domain-specific languages, and scientific workflows, e.g., Taverna [8] and Kepler [4], support the execution and replication of individual simulation experiments.Thereby, simulation models are part of the simulation data's provenance rather than being its primary subject. The development of a simulation model involves collecting and analyzing diverse data sources and executing various simulation experiments interleaved with the refinement, composition, or extension of

[*] The research was funded by the DFG (German Research Foundation) UH 66/18 "GrEASE" and by DFG CRC 1270 "Elaine".

the simulation model. As the generation of a simulation model is a highly intricate process, the accessibility of entities and diverse activities that contributed to its generation is as important as the accessibility of the simulation model itself. To capture the provenance of simulation models within and beyond individual simulation studies, we will exploit the PROV Data Model (PROV-DM) [2]. In combination with simulation experiments as first class entities and a multi-level approach, (nearly) the full tale behind a simulation model and its development can be revealed.

2 Exploiting PROV-DM for Simulation Model Development

The potential of PROV-DM in describing the provenance of a simulation model shall be illuminated based on a concrete biochemical model. The Wnt/β-catenin signaling is involved in central cellular processes, such as differentiation, proliferation, and migration of cells. As a central signaling pathway, significant efforts have been dedicated to understand the mechanisms of the pathway by developing a variety of simulation models. In [7], we presented a preliminary provenance model to relate a Wnt/β-catenin simulation model to earlier simulation models and data. This provenance model has been refined, as seen in Fig. 1, and transferred to PROV-DM. The connections of the five models (M1-M3') to other entities and activities are described in the following part.

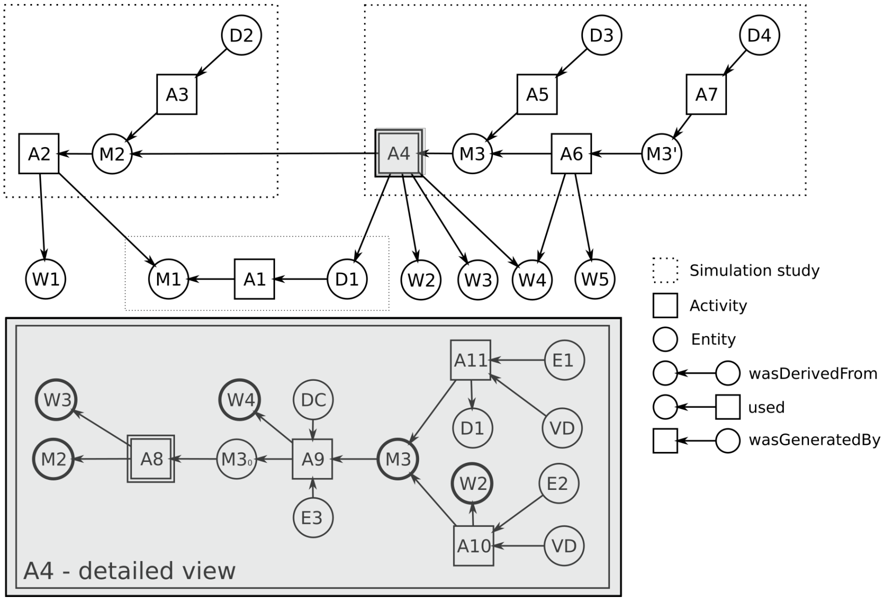

Fig. 1. Provenance model at multiple levels (in gray: detailed view of $A4$). The letters refer to: W - wet-lab data, D - model validation data, M - simulation model, A - activity like simulation experiment, DC - data calibration result, VD - model validation result, E - simulation experiments.

Data, hypotheses and model development: Data artifacts are used as input (*W3* - LRP6 initial values - in *A8*), for calibration (*W4* - nuclear β-catenin - in *A9*), and for validation (*D1* - cross validation with data produced with model *M1* - in *A11*). Assigning roles to the *used* relationship between activities and data artifacts facilitates assessing the diverse data sources and how they were used in generating simulation models.

Roles between activities and simulation models, such as *used* for adaptation (*M1* - for a different cell type - *A2*), extension, or *used* for composition (*M2* - by a membrane model - *A4, M3* - by a ROS model - *A6*), allow to assess the relationships between simulation models and to reuse entities and activities for a simulation model's progeny [6].

Simulation experiments: During the development of a simulation model, diverse simulation experiments, such as parameter scans, sensitivity analysis, simulation-based optimization, or statistical model-checking, are executed, alternating with phases of simulation model refinement, extension, or composition. Simulation experiments are part of a simulation model's generating process: directly, e.g., in terms of simulation-based optimization or parameter fitting (*E3*), or indirectly, by providing insights into the simulation model's behavior based on which the simulation model can be refined, extended, composed, or found to be valid (*E1, E2*). In addition, specifications of simulation experiments form important entities of a simulation model's provenance in their own right. They give substance to the generation process of a simulation model [7] and allow reusing simulation experiments across simulation models for consistency checks [6].

Activities at different levels: Similarly, as complex simulation models require to integrate description levels at multiple levels of abstractions, the "requirement of providing details at different levels of abstraction or from different viewpoints is (also) common in provenance systems" [5]. As the development of a simulation model is an intricate process, we cannot expect activities such as *A2, A4* or *A6* to be monolithic. For example, developing the simulation model *M3* relied on diverse simulation experiments which become visible by a more refined account (view) of activity *A4*. First, based on *M2* and wet-lab data, a model ($M3_0$) was derived (*A8*) which was subject to a calibration experiment (*A9, E3*) and later was validated by further simulation experiments (*A10, A11* and *E2, E1*, respectively), again based on different wet-lab data. Whereas those experiments and activities can be directly executed, the model ($M3_0$) itself has been composed of two simulation models which have been validated separately [6] and whose simulation experiments have been reused, which again would add a more fine grained account to the provenance model.

3 Towards an PROV Ontology for Simulation Model Development

A PROV ontology defines a specialization of PROV-DM. Our small case study already identified important ingredients of such an ontology: a) specific types of entities, e.g., data, theories, simulation experiments, and simulation models,

b) specific roles between specific types of entities, e.g., used as input, for calibration, for validation (between data and generation process), used for adaptation, extension, composition (between simulation models and generation process), c) specific refinement of activities: successive refinement of activities down to a level where simulation experiment specifications define activities and thus are ready to be executed, and d) specific inference strategies, e.g., warning if the same data have been used for calibration and validation, or validation experiments can be reused among descendants to check consistency. To approach a provenance ontology for simulation models, we are currently applying PROV-DM to additional simulation models in systems biology but also in other domains such as demography. In addition, we explore the potential of the provenance information for consistency checks by reusing simulation experiments across simulation models and studies.

References

1. Grimm, V., Polhill, P., Touza, J.: Documenting Social Simulation Models: The ODD Protocol as a Standard. In: Simulating Social Complexity - A handbook, pp. 117–133. Springer (2013)
2. Groth, P., Moreau, L.: Prov-overview. an overview of the prov family of documents (2013), https://www.w3.org/TR/prov-overview/
3. Hucka, M., Finney, A., Sauro, H.M., et al.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19**(4), 524–531 (2003)
4. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y.: Scientific workflow management and the kepler system. Concurrency and Computation: Practice and Experience **18**(10), 1039–1065 (2006)
5. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., et al.: The open provenance model core specification (v1. 1). Future generation computer systems **27**(6), 743–756 (2011)
6. Peng, D., Warnke, T., Haack, F., Uhrmacher, A.M.: Reusing simulation experiment specifications in developing models by successive composition – a case study of the wnt/β-catenin signaling pathway. Simulation: Transactions of the Society for Modeling and Simulation International **93**(8), 659–677 (April 2017)
7. Ruscheinski, A., Uhrmacher, A.M.: Provenance in modeling and simulation studies - bridging gaps. In: Winter Simulation Conference 2017. pp. 872–883. IEEE (2017)
8. Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., et al.: The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic acids research **41**(W1), W557–W561 (2013)