Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

CREATING PROV-DM GRAPHS FROM MODEL DATABASES

Pia Wilsdorf
Adelinde M. Uhrmacher

Institute for Visual and Analytic Computing
University of Rostock
Albert-Einstein-Str. 22
Rostock, 18059, GERMANY

ABSTRACT

Documenting the provenance of the main products of a simulation study plays a crucial role in improving the understanding of mechanistic, biological models as well as their reproducibility and credibility. With model databases already an ample collection of simulation models, including metainformation and source files, exists. In this paper, we bridge the gap between the information contained in model databases and the PROV-DM provenance standard, which allows making the diverse products and their relationships formally explicit. We present a procedure for creating PROV-DM graphs from model database entries, and illustrate the approach based on ten different models from the BioModels database. These case studies demonstrate the advantages of having a standardized provenance view in addition to the regular database entries, i.e., enhanced means for visualizing the structure of the simulation study and the curation process.

1 INTRODUCTION

Developing a valid simulation model to explain, analyze or predict real-world processes (e.g., of a biological system) is a complex and intricate task. Documenting the provenance of the main products of a simulation study plays a crucial role in improving the understanding of simulation studies as well as their reproducibility and credibility. Provenance, in general, refers to the "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" (Moreau and Groth 2013). In practice, the products of a simulation study come in the form of scripts, notebooks, or reports based on reporting guidelines that are shared via repositories or databases and accompanied by metainformation. Another way of expressing provenance is using the provenance standard PROV-DM, which provides a graph-based notation and thus enables better visualization and formal analysis of how products evolved. Although these two approaches to provenance are quite different, they can be regarded as complementary. Integrating these two views on provenance is an ongoing challenge.

A recent effort for combining provenance graphs with another form of model and study documentation is ODD+P (Reinhardt et al. 2018). There, protocols based on the reporting guideline ODD (Overview, Design concepts, Details) for agent-based models were enriched with provenance information in the Open Provenance Model. Another approach was the manual extraction of PROV-DM graphs from scientific publications to reveal the relations between a family of simulation models (Budde et al. 2021). In addition, the growing number of automatic capturing techniques emphasizes the demand for provenance documented in provenance standards like PROV-DM. These techniques include the automatic capturing of provenance graphs from scripts based on the structure of the code and executions logs (Murta et al. 2015) or user annotations (McPhillips et al. 2015), as well as collecting provenance as part of a scientific workflow (Altintas et al. 2006), or analyzing the structure of electronic lab notebooks (Schröder et al. 2022).

A substantial addition to the existing research would be the integration of PROV-DM with model repositories and databases. Model repositories and databases already contain an ample collection of source files and metainformation. Moreover, they are widely used in their respective communities, and thus integration with provenance graphs could boost reproducibility and credibility for a plethora of existing and future models. BioModels (Malik-Sheriff et al. 2020), for instance, is currently the largest database of models in systems biology, and the largest database of curated computational models overall.

In this paper, we bridge the gap between the information contained in model databases and the PROV-DM provenance standard, and present a procedure to make the diverse products of a simulation study and their relationships explicit in a graph. The concepts can be applied to any model database, however, here we will use the BioModels database as a case study. By bringing model databases and PROV-DM together we expect to gain additional insights into a model's development. Especially the relations between different models as well as the model iterations and curation history will be illuminated by this provenance view. This will have an impact on the understanding and reproducibility of models, as the involved sources can be more easily identified and consistency checks can be carried out. Adoption of our procedure by the different model databases would immediately bring these benefits to a large number of users.

The outline of the paper is as follows. In Section 2, we introduce the PROV-DM standard and its specialization for simulation studies. In Section 3, we present the case study for this paper, which we will use to illustrate and evaluate our concept. In Section 4, we analyze the relationship between PROV-DM and model databases and devise a detailed translation. Finally, we present the results of our case study in Section 5, and close the paper with a discussion in Section 6.

2 PROV-DM SPECIALIZATION FOR SIMULATION STUDIES

PROV-DM, the PROV data model, allows representing provenance information formally as a directed, acyclic graph (Belhajjame et al. 2013). It also provides an intuitive graphical representation and facilitates complex queries if stored in a graph database. The PROV-DM standard considers three types, i.e., entities, activities, and agents. Entities and activities are connected via relations, with *used* and *wasGeneratedBy* being the most frequently used relations. Agents can be associated with entities, activities, or other agents to indicate responsibilities. The general nature of these concepts allows for PROV-DM to be applied to and specialized for a wide range of fields, e.g., to analyze traffic in the internet of things (Sadineni et al. 2021).

To apply provenance to the products of modeling and simulation, a specialization of the PROV-DM types was developed. It identified the main entities to be Data, the Simulation Model, and Simulation Experiments, and whether those have been used by activities as input, for calibration or validation, used for adaptation, extension, or composition (Ruscheinski and Uhrmacher 2017; Ruscheinski et al. 2018). This specialization was further refined with simulation models of signaling pathways in mind (Budde et al. 2021) and based on discussions about the role of conceptual modeling in simulation studies (Wilsdorf et al. 2020). As a result, the entity types Research Question, Assumption, Requirement, Qualitative Model, Simulation Model, Simulation Experiment, Simulation Data, Wet-lab Data were proposed. In addition, four activity types were specified, i.e., Building Simulation Model, Calibrating Simulation Model, Validating Simulation Model, and Analyzing Simulation Model.

3 CASE STUDY: THE BIOMODELS DATABASE

Analyzing and understanding complete biological processes requires models to be available for reuse and composition by other researchers. BioModels is a platform that facilitates the sharing of FAIR (findable, accessible, interoperable, and reusable) quantitative models (Malik-Sheriff et al. 2020). The database is free and openly accessible at https://www.ebi.ac.uk/biomodels/.

Most models in BioModels are ordinary differential equation models, but recently also other types, e.g., logic-based or constraint-based models, are supported. Models have to be encoded in standardized formats such as SBML (Hucka et al. 2003) or CellML (Lloyd et al. 2004). From the model files, reaction

network diagrams can be generated to be also made available. Increasingly, also the simulation experiments are shared in separate formats, e.g., using SED-ML (Waltemath et al. 2011) or COPASI (Hoops et al. 2006). The community initiative Computational Modeling in Biology Network (COMBINE) coordinates the development of the various standards and their combination, e.g., to bundle all information needed to reproduce a simulation experiment, such as data, simulation model, or simulation experiment specification in an archive (Bergmann et al. 2014).

Models submitted to BioModels must adhere to the MIRIAM (Minimal Information Requested in the Annotation of Models (Novère et al. 2005)) reporting guidelines. MIRIAM requires to include at least a unique name, a citation associating the model to a publication, contact information for the model authors, the date and time of model creation and last modification, and the terms of distribution. Beyond that, to unambiguously identify the model components, models can be semantically annotated and linked to ontologies and other databases like the NCBI Taxonomy (Federhen 2011), the Gene Ontolgoy (Consortium 2004), or KEGG (Kanehisa et al. 2007). Furthermore, they can be cross-referenced with other models.

The models submitted to BioModels are independently curated to ensure that they are consistent with the referenced publication and that they produce the described numerical results. Over the past years, BioModels has become the world's largest repository of curated computational models. Currently (as of 21 March 2022), the database counts over 2300 published models, of which more than 1000 have been manually curated.

To demonstrate our concept, we create provenance graphs for ten different models from the BioModels database. We selected models with different submission dates (2008–2021) as over the years further specification formats and links to ontologies and other databases were established. We also chose models with different publication dates (1989–2021) as this also affects how detailed the documentation is.

Furthermore, we use the simulation study by Giordano et al. (BIOMD0000000955, last accessed 21 March 2022) as a running example to illustrate our concept in the following section. The objective of the simulation study was to build a model that predicts the course of the COVID-19 pandemic and the effects of various population-wide interventions. The model was built and tested based on real data of infections and deaths due to the SARS-CoV-2 virus in Italy. The results of that study showed that the COVID-19 pandemic can best be controlled with social-distancing measures and confirmed the benefit of widespread testing and contact tracing.

4 CREATING PROV-DM GRAPHS FROM A MODEL DATABASE

There are four main tasks when creating PROV-DM graphs from a model database: 1. recognizing the entities, 2. extracting the metainformation, 3. deriving the activities and relationships, and 4. connecting to related studies. In the following, we discuss these steps in detail.

Starting with the PROV-DM ontology described in Section 2, we are able to provide a translation to the provenance entities Simulation Model (SM), Simulation Experiment (SE), Simulation Data (SD), and Input Data (ID), and the provenance activities Building Simulation Model (BSM) and Analyzing Simulation Model (ASM). Since the curation of artifacts is an integral part of some model databases, such as BioModels, we introduce three additional types, i.e., the Curation Data (CD) entities, the Curating Simulation Model (C) activities, as well as the Publication (Pub) entities against which the uploaded artifacts are compared. Furthermore, we add the entity Conceptual Model (CM) to capture both the qualitative model as well as context information about the modeled system and the corresponding activity type Building Conceptual Model (CM). Other entities, such as research questions, assumptions, or requirements, can currently not be derived. Also, calibration and validation activities are not distinguishable from the more general "analyzing" experiments. Therefore additional, explicit annotations would be important.

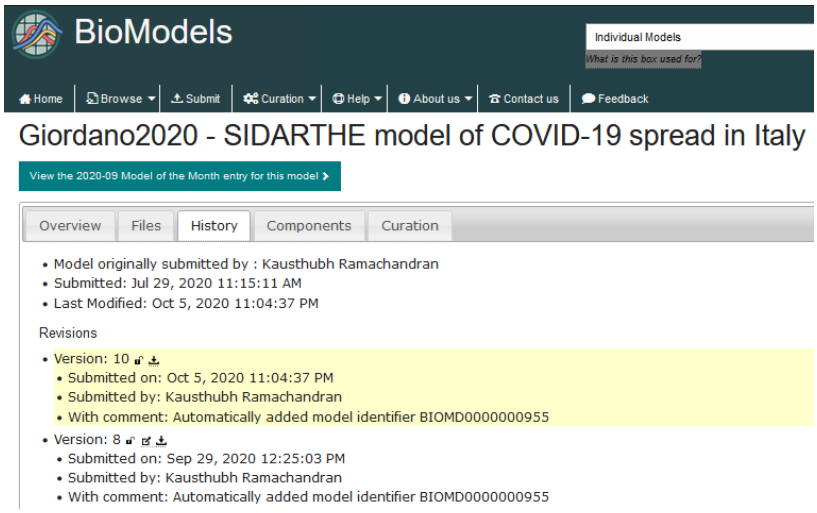

Figure 1: The BioModels page of the simulation study by Giordano et al. showing the most recent versions.

4.1 Recognizing the Entities

Each database entry contains valuable information that can be used to derive the nodes of the provenance graph, and to later fill them with metainformation and connect them via activities. In the BioModels database this information is, e.g., described in different tabs (i.e., Overview, Files, History, Components, and Curation as shown in Figure 1).

To identify all provenance entities, we have to go through the different model revisions one by one and analyze the information provided. Model databases usually show the latest public version of a simulation study, and previous revisions can be accessed via the menu. However, there may also exist private versions, i.e., versions only visible to the contributors. This is the case when the numbering from version 1 to n (current version) is not consecutive or marked as closed. We ignore the private versions in our provenance graphs as no metainformation is available for them. Only in the special cases, where there is no public version that can be used as input to the curation, private versions will be added as proxy entities within the provenance graph. For the simulation study by Giordano et al., there are four public model revisions available, for which several entities can be created (see Table 1). For each revision, at most one conceptual model entity can be added if ontology annotations exist that provide context about the simulation study, or if a file containing a conceptual diagram (e.g., a reaction network given as SVG) exists. Also, each revision must contain exactly one main simulation model entity. What format the simulation model may be specified in depends on the model database at hand. In the case of BioModels, the simulation models are usually given in SBML (Hucka et al. 2003) or CellML (Lloyd et al. 2004). Other entities can be recognized by their file extensions or formats as well. The simulation experiments in BioModels, for instance, are typically provided in COPASI (Hoops et al. 2006) or SED-ML (Waltemath et al. 2011) files. As a general rule, we create a new entity for each of such files found. However, we have to account for database-specific corner cases. For instance, in BioModels often both a COPASI and a SED-ML are provided for the same experiment. To clearly distinguish which files belong to which entity, the file names or descriptions can help, as they might contain hints such as “COPASI file of experiment xyz” or “SED-ML file of experiment xyz”. The entities of type simulation data and input data can be detected analogously, e.g., by finding CSV files containing raw data, or visualized data as PNGs.

Sometimes, the above-mentioned files are bundled and uploaded as an archive. In that case, the archive has to be extracted first before the files can be analyzed. E.g., COMBINE archives encoded in the OMEX format (Bergmann et al. 2014) are frequently used in the context of BioModels.

Moreover, the entities referring to the curation have to be added. If curation information exists, we can create a curation data entity (curation output) representing the figures that were reproduced, and an

Table 1: The provenance entities recognized for the different versions of the model by Giordano et al. CM–Conceptual Model, SM–Simulation Model, SE–Simulation Experiment, CD–Curation Data, Pub–Publication.

Version	Entities	Explanation
3	SM3, SE3	Initial upload of simulation model and simulation experiment
5	SM5, SE5	Minor update
8	SM8, SE8, CD8, Pub	Minor update and independent curation
10	CM10, SM10, SE10	Add context information and reproducible files

additional entity for the reference publication (curation input) against which the model is curated. Note that, e.g., in BioModels, currently for each database entry only one curation is shown, even if multiple ones were attempted. Either way, we have to find out based on which revision the curation was carried out, in order to create and relate the entities to the correct revision. If this mapping is not directly accessible, it can be accomplished by comparing the timestamps of the curations with the timestamps of the different revisions. For the Giordano et al. simulation study, we can derive that the only curation was carried out based on SM8, and thus the curation data and publication are added to version 8 (see Table 1).

4.2 Extracting the Metainformation

The result of the previous step is a set of entities. These entities can then be refined with metainformation. In the following, we will go through the different provenance entity types and discuss what metainformation can be extracted. In Table 2, we list the different entity types and their attributes, and provide an exemplary mapping to the current release of BioModels (as of March 2022), and give examples from the simulation study by Giordano et al. For instance, *Overview/hasTaxon* refers to the field named *hasTaxon* in the tab named *Overview* of the BioModels database, which may be linked to the concept *Homo Sapiens* of the NCBI Taxonomy (Federhen 2011). Apart from the listed attributes, each entity is also assigned an entity name (derived from the entity type and the version number) and a study name.

Simulation Model Relevant information for the simulation model entities are, first of all, a database-specific identifier and a short description which might include the abstract of the corresponding scientific publication and also information on how to reproduce the results. These are usually available for all models and easily accessible. Next, the list of variables, the list of initial values, and the list of parameters have to be added to the simulation model entity. These are not always made explicit in the database, and therefore can be left blank or extracted directly from the model specification. In the simulation model of Giordano et al., e.g., the different variables referring to the infection status (Susceptible, Infected, Diagnosed, Ailing, Threatened, and Healed) are extracted and assigned their initial values. If made explicit, also the modeling approach can be extracted, e.g., it could have been annotated using an ontology such as the Mathematical Modeling Ontology (MAMO 2022). Next, a reference to the simulation model can be easily extracted as well as the specification format (in BioModels usually SBML or CellML). Sometimes, various alternative formats of the model may be found in the same entry and added to the references list. For BioModels, these include, e.g., the OWL-based BioPAX exchange format (Demir et al. 2010) or Scilab (Campbell et al. 2010), which can all be autogenerated by the Systems Biology Format Converter (Rodriguez et al. 2016) from an SBML specification (often done in older database entries). Finally, we add some metainformation referring to the revision history to the simulation model entities. In particular, each public version is assigned a timestamp and the name of the submitter. E.g., the latest submission of the model by Giordano et al. was uploaded on October 5th, 2020. Another possibility would be to add the submitter as a PROV-DM agent, but we decided to make the agents not explicit in the provenance graph as this information can be easily

integrated in the simulation models and currently no other agent types exist. However, if further roles are added in the future, e.g., if the name of the curator would be made public, one might reassess this decision.

Conceptual Model We interpret the role of the conceptual model to store qualitative information (i.e., the qualitative model given as diagram or sketch), and also to provide context about the modeled system. The context can be annotated by using various ontologies depending on the database and also on the application domain. For example, from the BioModels database we can extract which organism (NCBI Taxonomy (Federhen 2011)), which biological processes (e.g., Gene Ontology (Consortium 2004)), or which disease was modeled (Human Disease Ontology (Schriml et al. 2021)), and based on data of which cell line (BRENDA Tissue Ontology (Gremse et al. 2010)). The qualitative model (in BioModels the reaction diagram showing all the participating model species and the types of reactions between them) can typically be recognized by its file type (PNG, SVG or verbal description in PDF). The URLs to this file can be extracted and added as a reference to the respective conceptual model entity, and their format can be added as well.

Simulation Experiment, Simulation Data, Input Data References, the file formats, and possibly descriptions can be extracted and added to the recognized simulation experiment, input data or simulation data entities, in a similar fashion to the extraction of the simulation model and the conceptual model. For example, for each version of the Giordano et al. simulation study, an SBML file is discovered to be part of the simulation model, and a COPASI and a SED-ML file as part of the simulation experiment. What formats are available depends on the database and what formats are currently supported or rather were supported at the time the model was uploaded.

Curation Data Metainformation of the curation data includes a short description, the software version used for simulation or plotting, the date and time of the curation, as well as the format and reference of the reproduced figures. E.g., for the Giordano et al. model Figures 2b, 2d, 3b, 3d, 4b, and 4d were reproduced using COPASI 4.27 (Build 217). The information is typically added to the database by the curator in a short comment. In the future, this comment could be further expanded to explicitly annotate which parameter values were required to successfully reproduce the data or figures from the publication.

Publication Finally, with respect to the curation, we also want to add information about the publication (i.e., reference to a journal article) to which the simulation results are compared. This information is usually given as a URL or DOI in the database entry.

4.3 Deriving the Activities and Relationships

Once the entities have been identified and filled with meta information, the activities can be created and relationship arrows can be drawn. Table 3 provides an overview of the different activity types and the types of entities they use or generate. Version by version, the available entities are taken into account to derive the necessary activity type as follows.

1. If a conceptual model entity exists, a Building Conceptual Model activity is created, with the conceptual model as its output, and the preceding conceptual model (if available) as input.
2. If a simulation model entity exists, a Building Simulation Model activity can be created, using the previous simulation model and conceptual model (if available) as input, and the new simulation model as output.
3. For each simulation experiment entity that exists, an Analyzing Simulation Model activity is added, with the simulation model and possibly data as input, and the simulation experiment, and (if available) corresponding simulation data as output.
4. If curation data and publication exist for the current version, a Curating Simulation Model activity is created, taking the publication as input, as well as the simulation model and possibly a simulation experiment and input or simulation data, and generating the curation data.
5. If the curation immediately produced a new version, add all entities of the following version to the outputs of the curation and skip the next version.

Table 2: The central provenance entity types with their attributes in simulation studies and their correspondence in the BioModels database. The rightmost column provides examples from the simulation study by Giordano et al. CM–Conceptual Model, SM–Simulation Model, SE–Simulation Experiment, SD–Simulation Data, ID–Input Data, CD–Curation Data, Pub–Publication.

Entity	**Attribute**	**BioModels**	**Example**
CM	Organism	Overview/hasTaxon	Homo sapiens
	Cell Line	Overview/occursIn	—
	Virus	Overview/hasTaxon	SARS-CoV-2
	Disease	Overview/hasProperty	COVID-19
	Biological Processes	Overview/isVersionOf, isHomologTo	Infectious Disease Pandemic
	Format	Files/Additional files	—
	Reference	Files/Additional files	—
SM	Identifier	Overview/Model Identifier	BIOMD0000000955
	Description	Overview/Short description	... We propose a new model that predicts the course of the epidemic to help plan...
	Variables	Components/Species/Species	Susceptible, Infected, Diagnosed, Ailing, Threatened, Healed
	Initial Values	Components/Species/Initial Concentration/Amount	0.9999963, 3.33333333E-6, 3.33333333E-7, 1.66666666E-8, ...
	Parameters	Components/Reactions/Parameters	zeta, kappa, ...
	Parameter Values	Components/Reactions/Parameters	0.125 1/d, 0.017 1/d, ...
	Modeling Approach	Overview/hasProperty, Overview/Modelling Approach(es)	Population Model
	Format	Overview/Format, Files/Model files	SBML (L3V1)
	Reference	Files/Model files	Giordano2020.xml
	Submission Date	History/Submitted on	Oct 5, 2020 11:04:37 PM
	Submitted By	History/Submitted by	Kausthubh Ramachandran
SE	Description	Files/Description	—
	Format	Files/Additional Files	COPASI, SED-ML
	Reference	Files/Additional files	Giordano2020.cps, Giordano2020.sedml
SD	Description	Files/Description	—
	Format	Files/Additional files	—
	Reference	Files/Additional files	—
ID	Description	Files/Description	—
	Format	Files/Additional files	—
	Reference	Files/Additional files	—
CD	Description	Curation/Curator's comment	...To reproduce Fig 2b, set Event_trigger_Fig3b = 0 ...
	Software	Curation/Curator's comment	COPASI 4.27 (Build 217)
	Curation Date	Curation/updated	05 Oct 2020, 23:03:41
	Format	Curation/Figure, Curator's comment	PNG
	Reference	Curation/Figure	Figures 2b, 2d, 3b, 3d, 4b, 4d
Pub	Reference	Overview/Related Publication	https://doi.org/10.1371/journal.pcbi.1002437

Table 3: Types of provenance activities, what inputs they use, and what outputs they generate.

Activity	**Used**	**Generated**
Building Conceptual Model	Conceptual Model	Conceptual Model
Building Simulation Model	Conceptual Model, Simulation Model	Simulation Model
Analyzing Simulation Model	Simulation Model, Input Data	Simulation Experiment, Simulation Data
Curating Simulation Model	Publication, Simulation Model, Simulation Experiment, Input Data, Simulation Data	Curation Data, Simulation Model, Simulation Experiment, Conceptual Model

The latter can be again evaluated by comparing the timestamp of the revisions and curations. If the difference between the two dates lies approximately within a day, we can assume a causal relationship between the curation process and the creation of new (functional and reproducible) versions of the entities. E.g., in the simulation study by Giordano et al., the entities of CM10, SM10 and SE10 were created as part of the curation activity C8, see Figure 2.

4.4 Connecting to Related Studies

Finally, the database page may provide links to previous models based on which the model at hand was developed. In BioModels, e.g., the related models are referenced via the *isDerivedFrom*-qualifier. These relationships are of particular interest when creating provenance graphs to tell the tale about a family of models (Budde et al. 2021). Depending on whether a database entry exists for the related model or only the publication is referenced, either the same procedure as described above is then applied for the related studies recursively, or a single entity is added as proxy for the related model. To connect two studies, a used-relationship is drawn from the first Building Simulation Model activity of the current study to the last version of the related model (or the related model proxy). In the case of our running example, however, no information about related model is given.

5 RESULTS

We applied our procedure to ten different models from the BioModels database. The provenance graphs of these case studies are available in a Git repository as SVG and PDF, accompanied by separate documents containing the metainformation of the entity and activity nodes. Starting from these ten studies and continuing with related simulation studies, we received overall 26 provenance graphs including 143 entities and 101 activities. The most frequent entities we encountered in the case studies were the Simulation Model (55 entities) and the Simulation Experiment (23 entities). Curation Data and Publication entities by construction occurred at most once per study (20 entities each). The entities Conceptual Model (15 entities) and Simulation Data (ten entities) were less common, and no input data (0 entities) was found in the examples.

Figure 2 shows the provenance graph of our running example, i.e., the COVID-19 simulation study by Giordano et al. (2020). It shows that the simulation model was updated and associated with a simulation experiment three times before it finally was independently curated (C8) against the publication. The curation generated new versions of the simulation model and the simulation experiment, with which the curator was able to reproduce the figures contained in the paper (figures referenced in the curation data entity). Furthermore, a conceptual model (CM10) was added during the curation, where the model was annotated with context information about the modeled organism, virus, and disease.

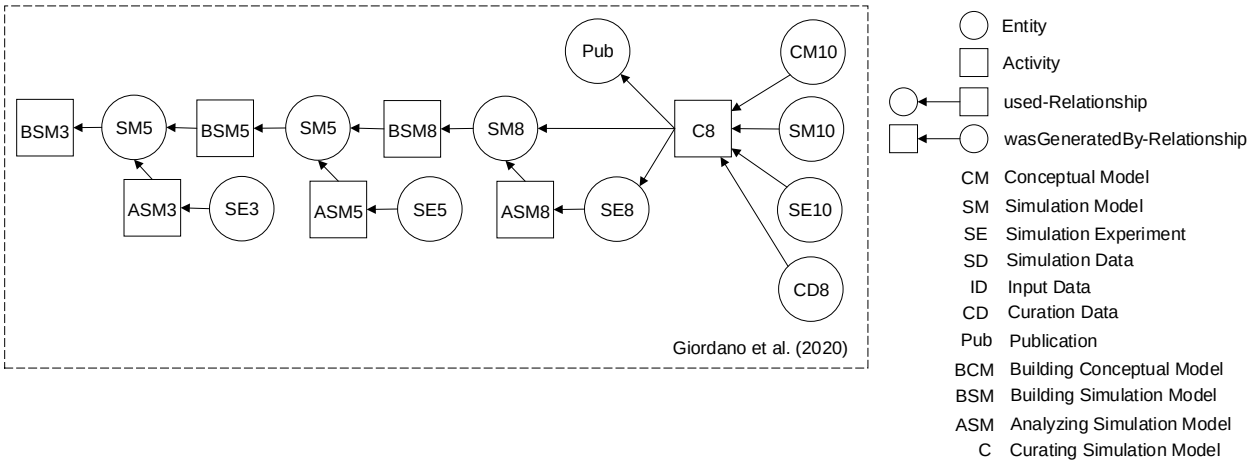

Figure 2: Provenance graph generated from the BioModels entry BIOMD0000000955.

A more complex example can be seen in Figure 3. The provenance graph visualizes the curation process of the simulation study by Padala et al. (2017) and its related simulation studies. The simulation study investigated the dynamics of signaling responses in the ERK, PI3K/Akt, and Wnt/β-catenin signaling network during genetic and epigenetic changes in cancer. The results show how malfunctions of network components lead to an additive effect on cancer growth. After uploading the simulation model to BioModels, it was curated (C1) against the publication, which resulted in an auto-generated reaction network as part of the conceptual model (CM2), an updated simulation model (SM2), and curation data (CD1). The conceptual model is later updated (BCM3) with various context information, and the simulation model is enriched (BSM3) with information about the initial concentrations and parameter values. By following the relationship-arrows back in time, we see that the model was built on three related models. These models are again related to other models, e.g., we see that both Padala et al. and Orton et al. were built on the model by Brown et al. and that the model by Kim et al. was built on two other models (Lee et al. 2003 and Cho et al. 2003) for which, however, no BioModels entries exist.

6 DISCUSSION

In our concept and the case studies, we have demonstrated how provenance graphs in PROV-DM can be a valuable addition to model databases, such as the BioModels database, to make the development and curation process transparent. The provenance graphs visualize the various entities and activities that played a role in producing the study results and reveal the dependencies between related studies. Currently, not all types of entities (such as Simulation Data or Input Data) are published for all models, however, we believe that with the increasing awareness for reproducibility these will become more common. For larger simulation studies it could also be interesting to add agents and roles to the provenance graphs to make explicit who contributed during modeling, analysis, or curation. Naturally, the provenance graphs we generate from model databases focus mostly on the curation history of a model, as models are typically uploaded to the database after they have been published in a journal. To get a more complete picture of the simulation study, the curation history could be integrated with the development history, which could be extracted, e.g., from GitHub commit logs (Packer et al. 2019).

The created provenance view will help scientists to understand the simulation studies better, and might also enable them to uncover mistakes as well as unclear or missing documentation. E.g., the provenance representation might help to identify if or when a curation needs to be repeated after updates on the central entities, i.e., simulation models or simulation experiments. Furthermore, it might allow users to trace back why models do not produce the same results as in the publication, ultimately improving the quality and reproducibility of studies.

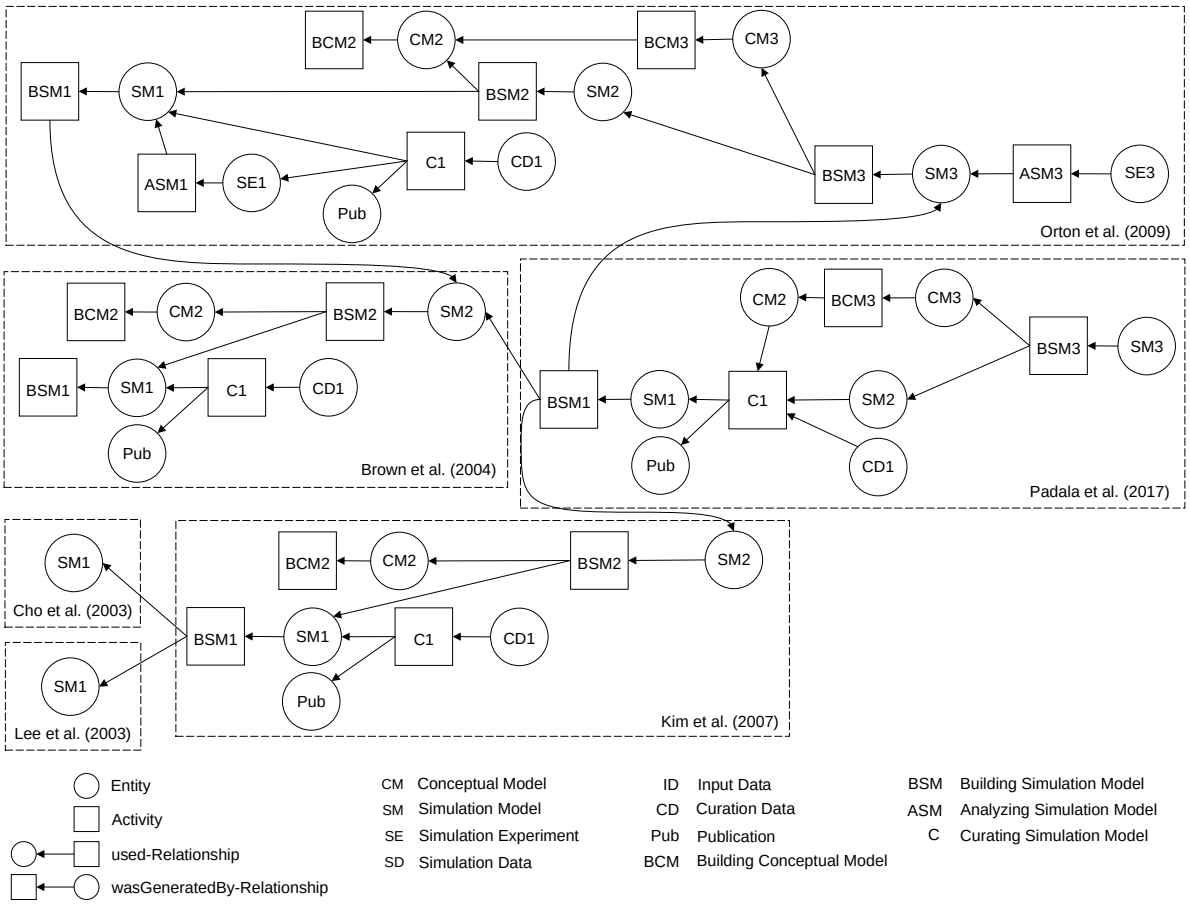

Figure 3: Provenance graph created from the BioModels entry BIOMD0000000652.

For now we have created the provenance graphs by hand as proof of concept. In the future, however, the presented approach can be implemented as an integral part of the various model databases to automatically create provenance graphs for submitted models, and therefore not to burden the modelers with this task. Also, drawing sophisticated conclusions from the graphs can be supported by tools that automatically interpret the changes between different versions of the simulation models (Scharm et al. 2015) or other entities. But correctly interpreting changes in computational models is a difficult problem and will need to be investigated further (Scharm et al. 2018).

Although we illustrated and tested our concept specifically for the BioModels database, the general procedure can be transferred to other model databases, e.g., the CoMSES Model Library for computational models in social and ecological sciences (Janssen et al. 2008). The CoMSES Model Library offers features that are key to our approach, including model revisions, peer review, file uploads, and annotations. However, there are some fundamental differences between BioModels and CoMSES that would need to be addressed. First of all, the scope of BioModels is reserved for models from systems biology, typically specified as ODE systems. This well-defined domain provides numerous ontologies and standardized formats that can be exploited to extract provenance entities and metainformation. CoMSES on the other hand focuses on agent-based models but is open to a wide range of application fields. While models can be tagged with keywords, ontologies or structured vocabularies rarely exist. Furthermore, in agent-based modeling frameworks, such as NetLogo (Tisue and Wilensky 2004), often no clear separation of concerns between the simulation model and simulation experiment exists, which limits the types of provenance entities and activities that can be created. However, with wider adoption of specification languages and frameworks

for reproducible simulation experiments like the NLRX R package (Salecker et al. 2019) this gap can be overcome. Finally, details about the various entities and activities are frequently described in a supplementary document based on a reporting guideline such as ODD (Grimm et al. 2020). We plan to explore the relationship between these verbal narrative-based documentations and PROV-DM in future work.

ACKNOWLEDGMENTS

This work was funded by the research grant DFG UH-66/18 'GrEASE'.

REFERENCES

Altintas, I., O. Barney, and E. Jaeger-Frank. 2006. "Provenance Collection Support in the Kepler Scientific Workflow System". In *Provenance and Annotation of Data*, edited by L. Moreau and I. Foster, 118–132. Berlin, Heidelberg: Springer Berlin Heidelberg.

Belhajjame, K., R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker et al. 2013. "PROV-DM: the PROV Data Model". *W3C Recommendation*.

Bergmann, F. T., R. Adams, S. Moodie, J. Cooper, M. Glont, M. Golebiewski, M. Hucka, C. Laibe, A. K. Miller, D. P. Nickerson et al. 2014. "COMBINE Archive and OMEX Format: One File to Share all Information to Reproduce a Modeling Project". *BMC Bioinformatics* 15(1):1–9.

Budde, K., J. Smith, P. Wilsdorf, F. Haack, and A. M. Uhrmacher. 2021. "Relating Simulation Studies by Provenance—Developing a Family of Wnt Signaling Models". *PLOS Computational Biology* 17(8):1–27.

Campbell, S. L., J.-P. Chancelier, and R. Nikoukhah. 2010. "Modeling and Simulation in SCILAB". In *Modeling and Simulation in Scilab/Scicos with ScicosLab 4.4*, 73–106. Springer.

Consortium, G. O. 2004, 01. "The Gene Ontology (GO) Database and Informatics Resource". *Nucleic Acids Research* 32(suppl_1):D258–D261.

Demir, E., M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'eustachio, C. Schaefer, J. Luciano et al. 2010. "The BioPAX Community Standard for Pathway Data Sharing". *Nature Biotechnology* 28(9):935–942.

Federhen, S. 2011, 12. "The NCBI Taxonomy database". *Nucleic Acids Research* 40(D1):D136–D143.

Giordano, G., F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, and M. Colaneri. 2020. "Modelling the COVID-19 Epidemic and Implementation of Population-wide Interventions in Italy". *Nature Medicine* 26(6):855–860.

Gremse, M., A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg. 2010, 10. "The BRENDA Tissue Ontology (BTO): The first All-integrating Ontology of all Organisms for Enzyme Sources". *Nucleic Acids Research* 39(suppl_1):D507–D513.

Grimm, V., S. F. Railsback, C. E. Vincenot, U. Berger, C. Gallagher, D. L. DeAngelis, B. Edmonds, J. Ge, J. Giske, J. Groeneveld, A. S. A. Johnston, A. Milles, J. Nabe-Nielsen, J. G. Polhill, V. Radchuk, M.-S. Rohwäder, R. A. Stillman, J. C. Thiele, and D. Ayllón. 2020. "The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism". *Journal of Artificial Societies and Social Simulation* 23(2):7.

Hoops, S., S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. 2006. "COPASI—a COmplex PAthway SImulator". *Bioinformatics* 22(24):3067–3074.

Hucka, M., A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov et al. 2003. "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models". *Bioinformatics* 19(4):524–531.

Janssen, M. A., L. N. Alessa, M. Barton, S. Bergin, and A. Lee. 2008. "Towards a Community Framework for Agent-based Modelling". *Journal of Artificial Societies and Social Simulation* 11(2):6.

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. 2007, 12. "KEGG for Linking Genomes to Life and the Environment". *Nucleic Acids Research* 36(suppl_1):D480–D484.

Lloyd, C. M., M. D. Halstead, and P. F. Nielsen. 2004. "CellML: Its Future, Present and Past". *Progress in Biophysics and Molecular Biology* 85(2):433–450. Modelling Cellular and Tissue Function.

Malik-Sheriff, R. S., M. Glont, T. V. N. Nguyen, K. Tiwari, M. G. Roberts, A. Xavier, M. T. Vu, J. Men, M. Maire, S. Kananathan, E. L. Fairbanks, J. P. Meyer et al. 2020. "BioModels — 15 Years of Sharing Computational Models in Life Science". *Nucleic Acids Research* 48(D1):D407–D415.

MAMO 2022. Mathematical Modelling Ontology. https://bioportal.bioontology.org/ontologies/MAMO, accessed 16^{th} March.

McPhillips, T. M., T. Song, T. Kolisnik, S. Aulenbach, K. Belhajjame, K. Bocinsky, Y. Cao, F. Chirigati, S. C. Dey, J. Freire, D. N. Huntzinger, C. Jones, D. Koop, P. Missier, M. Schildhauer, C. R. Schwalm, Y. Wei, J. Cheney, M. Bieda, and B. Ludäscher. 2015. "YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts". *CoRR* abs/1502.02403.

Moreau, L., and P. Groth. 2013. "Provenance: An Introduction to PROV". *Synthesis Lectures on the Semantic Web: Theory and Technology* 3(4):1–129.

Murta, L., V. Braganholo, F. Chirigati, D. Koop, and J. Freire. 2015. "noWorkflow: Capturing and Analyzing Provenance of Scripts". In *Provenance and Annotation of Data and Processes*, edited by B. Ludäscher and B. Plale, 71–83. Cham: Springer International Publishing.

Novère, N. L., A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes et al. 2005. "Minimum Information Requested in the Annotation of Biochemical Models (MIRIAM)". *Nature Biotechnology* 23(12):1509–1515.

Packer, H. S., A. Chapman, and L. Carr. 2019. "GitHub2PROV: Provenance for Supporting Software Project Management". In *Proceedings of the 11th International Workshop on Theory and Practice of Provenance*, edited by T. Moyer and S. Roy. Philadelphia, PA: USENIX Association.

Padala, R. R., R. Karnawat, S. B. Viswanathan, A. V. Thakkar, and A. B. Das. 2017. "Cancerous Perturbations within the ERK, PI3K/Akt, and Wnt/β-catenin Signaling Network Constitutively Activate Inter-pathway Positive Feedback Loops". *Molecular BioSystems* 13:830–840.

Reinhardt, O., A. Rucheinski, and A. M. Uhrmacher. 2018. "ODD+P: Complementing the ODD Protocol with Provenance Information". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 727–738. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Rodriguez, N., J.-B. Pettit, P. Dalle Pezze, L. Li, A. Henry, M. P. van Iersel, G. Jalowicki, M. Kutmon, K. N. Natarajan, D. Tolnay et al. 2016. "The Systems Biology Format Converter". *BMC Bioinformatics* 17(1):1–7.

Ruscheinski, A., D. Gjorgevikj, M. Dombrowsky, K. Budde, and A. M. Uhrmacher. 2018. "Towards a PROV Ontology for Simulation Models". In *Provenance and Annotation of Data and Processes*, 192–195.

Ruscheinski, A., and A. M. Uhrmacher. 2017. "Provenance in Modeling and Simulation Studies - Bridging Gaps". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 872–883. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sadineni, L., E. S. Pilli, and R. B. Battula. 2021. "Ready-IoT: A Novel Forensic Readiness Model for Internet of Things". In: *Proceedings of the IEEE 7th World Forum on Internet of Things*, June 14th-31st, New Orleans, LA, USA, 89-94.

Salecker, J., M. Sciaini, K. M. Meyer, and K. Wiegand. 2019. "The NLRX R Package: A next-Generation Framework for Reproducible NetLogo Model Analyses". *Methods in Ecology and Evolution* 10(11):1854–1863.

Scharm, M., T. Gebhardt, V. Touré, A. Bagnacani, A. Salehzadeh-Yazdi, O. Wolkenhauer, and D. Waltemath. 2018. "Evolution of Computational Models in BioModels Database and the Physiome Model Repository". *BMC Systems Biology* 12(1):1–10.

Scharm, M., O. Wolkenhauer, and D. Waltemath. 2015. "An Algorithm to Detect and Communicate the Differences in Computational Models Describing Biological Systems". *Bioinformatics* 32(4):563–570.

Schriml, L. M., J. B. Munro, M. Schor, D. Olley, C. McCracken, V. Felix, J. Baron, R. Jackson, S. Bello, C. Bearer, R. Lichenstein, K. Bisordi, N. C. Dialo, M. Giglio, and C. Greene. 2021, 11. "The Human Disease Ontology 2022 update". *Nucleic Acids Research* 50(D1):D1255–D1261.

Schröder, M., S. Staehlke, P. Groth, J. B. Nebe, S. Spors, and F. Krüger. 2022. "Structure-based Knowledge Acquisition from Electronic Lab Notebooks for Research Data Provenance Documentation". *Journal of Biomedical Semantics* 13(1):1–22.

Tisue, S., and U. Wilensky. 2004. "Netlogo: A Simple Environment for Modeling Complexity". In *Proceedings of the Fifth International Conference on Complex Systems*, edited by A. Minai and Y. Bar-Yam, 16–21: Boston, MA.

Waltemath, D., R. Adams, F. T. Bergmann, M. Hucka, F. Kolpakov, A. K. Miller, I. I. Moraru, D. Nickerson, S. Sahle, J. L. Snoep et al. 2011. "Reproducible Computational Biology Experiments with SED-ML-the Simulation Experiment Description Markup Language". *BMC Systems Biology* 5(1):198.

Wilsdorf, P., F. Haack, and A. M. Uhrmacher. 2020. "Conceptual Models in Simulation Studies: Making it Explicit". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 2353–2364. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

PIA WILSDORF is a Ph.D. candidate in the Modeling and Simulation group at the University of Rostock. Her e-mail address is pia.wilsdorf@uni-rostock.de.

ADELINDE M. UHRMACHER is a Professor at the Institute for Visual and Analytic Computing, University of Rostock, and head of the Modeling and Simulation group. Her e-mail address is adelinde.uhrmacher@uni-rostock.de.