Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

ODD+P: COMPLEMENTING THE ODD PROTOCOL WITH PROVENANCE INFORMATION

Oliver Reinhardt
Andreas Rucheinski
Adelinde M. Uhrmacher

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

ABSTRACT

Provenance information about a simulation model is information about people, artifacts, and processes that have contributed to its generation. It increases trust into the quality and validity of simulation models. Model documentation standards such as the ODD protocol have a similar goal, but are mostly concerned with "what has been generated", and less with "how it has been generated". Complementing ODD with provenance information offers a more structured approach to the "what" and fills the gap regarding the "how". Thereby, simulation experiments play a crucial role, and are treated as first class artifacts, as are simulation models, data sources, and theories. The compliance to the Open Provenance Model allows using established tools for inferring the model's origin. The approach is of particular value for models that are based on various data sources, theories, and earlier models, as we will show based on a model about migration from Senegal to Europe.

1 INTRODUCTION

Provenance refers to information about how a product has been generated (Simmhan et al. 2005). 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"(Groth and Moreau 2013).

Applying provenance in modeling and simulation requires identifying central processes and products of modeling and simulation and putting them into relation (Ruscheinski and Uhrmacher 2017). The main products of simulation studies are the data produced and the simulation model itself. The provenance of data, may it be generated in-silico, in-vitro or in-vivo, has been the subject of major research efforts during the last two decades and accordingly a diversity of software tools and platforms are available, such as Fairdom (Wolstencroft et al. 2017) or VisTrails (Callahan et al. 2006), which provide a rich portfolio of methods to replicate or reproduce data. In these efforts, simulation models form a part of simulation data provenance. How the simulation models themselves have been generated has received little attention yet, with a few exceptions (Ruscheinski and Uhrmacher 2017). Although efforts have been dedicated to making simulation models accessible and facilitating their reuse, such as the ODD protocol (Overview, Design concepts, Details; Grimm et al. 2010), or the Preferred Model Reporting Requirements (PMRR; Rahmandad and Sterman 2012), these focus on the product, i.e., what the model looks like, rather than the process, i.e., how the model has been generated.

To address this deficiency, we propose to complement the documentation of simulation models with a provenance model about how a simulation model has been generated. We concentrate our efforts on ODD, a standard for documenting agent-based simulation models, and the Open Provenance Model (OPM; Moreau et al. 2011), as the standard for provenance description. We will begin with an introduction into

ODD and OPM to afterward describe our approach, which we coin ODD+P (ODD + Provenance) and apply it to a simulation model about migration from Senegal to Europe.

2 The ODD Protocol

One of the main artifacts of interest in simulation studies is the simulation model. Detailed information about the model is crucial for assessing its validity, and for reproducing and reusing it. The ODD (Overview, Design concepts, and Details) protocol (Grimm et al. 2006) defines a structure to describe agent-based and individual-based simulation models. It makes the documentation of such models more rigorous, by defining the necessary elements for decribing a simulation model), and more accessible, by prescribing how these elements shall be organized. In recent years ODD has been widely adopted for agent-based modeling (Grimm et al. 2010; Schulze et al. 2017). The OpenABM computational model library recommends ODD for documenting the uploaded models (CoMSES Network 2018). ODD demands seven elements of model description (see Table 1), organized into three blocks: *Overview*, *Design concepts*, and *Details*.

Table 1: The seven elements of the ODD protocol (Grimm et al. 2010).

Overview	1. Purpose 2. Entities, state variables, and scales 3. Process overview and scheduling
Design concepts	4. Design concepts (basic principles, emergence, adaptation, objectives, learning, prediction, sensing, interaction, stochasticity, collectives, observation)
Details	5. Initialization 6. Input data 7. Submodels

In the first block, *Overview*, the model is put into a context by describing its *Purpose* (first element), general information about the model structure (*Entities, state variables, and scales*), and the modeled processes (*Process overview and scheduling*). The third part also entails a description of how the model deals with time (discrete or continuous time) and how events are scheduled. All in all this shall allow an experienced reader to relate the model to other models of the field and assess its overall design and complexity. In the second block, *Design concepts*, the principles of the model's design are discussed. The block consist of only one element, also called *Design concepts*, which can consist of several aspects common in agent-based models. This includes information, e.g., on whether the agents are stochastic, how agents interact with each other, what specific sensing mechanisms are used. Much of the information in this block is not needed to replicate the model, but does inform the reader about the design decision made.

Our focus is on the third and final block, *Details*, which gives information necessary to re-implement the model. This includes model initialization, input data, and the model's submodels. In the *Initialization* element, the state of the model entities at the very beginning of a simulation run shall be laid out. How many entities of each type are there? How are they attributed? If the model is spatial, how are the agents distributed in space? And if it contains a social network, how are agents linked with each other? This also includes references to the data used to generate these initial conditions. In the *Input data* element the role and source of external data shall be described. Finally, in *Submodels* all the model components that deal with the processes introduced in the Overview block shall be explained in detail. This shall include "appropriate levels of explanation an justification", and at the same time remain concise and readable (Grimm et al. 2010). In case submodels are derived from independently published models or theories, the ODD contains references to the relevant literature.

While the widespread adoption of ODD shows it to be viewed as both practical and beneficial, the protocol has some shortcomings. An ODD document only describes a single model version, the documentation

of the evolution of the model is not foreseen by the protocol. Grimm et al. 2006 recommend a separate ODD description for each published model version, until a better solution has been established. While the protocol provides a general structure for documentation, it does not propose a structure for the ODD-elements themselves. This is especially problematic for more complex elements, such as *Submodels*, where entire models have to be described. As a consequence, those descriptions are often not well structured (Grimm et al. 2010). While the authors of ODD find that the model documentation should be supplemented with a documentation of simulation experiments conducted with the model (Grimm et al. 2010), this is not addressed by the standard.

3 The Open Provenance Model

The Open Provenance Model (OPM; Moreau et al. 2011) allows to describe provenance information as a directed graph, where nodes represent *artifacts*, *processes*, or *agents*, while edges indicate dependencies. Here, *artifacts* are digital representations of entities within a computer system, in our case component models, data sources or experiment specifications. *Processes* represent activities performed with artifacts to generate new artifacts. And *agents* are the entities enabling and controlling the processes. Between these elements five dependencies are distinguished:

1. an artifact was *used* in a process,
2. an artifact *was generated by* a process,
3. a process *was controlled by* an agent,
4. a process *was triggered by* another process, and
5. an artifact *was derived from* another artifact.

In this paper we will focus on artifacts and processes, ignoring the agents, as we are less interested in "who did what" than in the "how was it done".

4 COMPLEMENTING ODD WITH PROVENANCE INFORMATION

A first step is to identify artifacts and processes of modeling and simulation. In (Ruscheinski and Uhrmacher 2017) we applied the Open Provenance Model to a cell biological model, and identified simulation model, simulation experiment, and data, which can be used as input, for validation, and for calibration, as artifacts of the simulation process. In other cases established theories and models also contribute to the simulation model. In the following, we will focus on these types of artifacts:

1. Data sources, from which data is used as an input in the sense of the ODD *Input* element, as well as for calibration and validation. The provenance model explicitly shows the different data sources and their role, allowing for reasoning about their relation.
2. Models and theories from literature, which form the theoretical foundation of the model. This includes, but is not limited to, theories, e.g., behavioral theories in models from the social sciences, simulation models, or statistical models, which might form submodels of the developed simulation model, or upon which the simulation model might be based. This allows for the assessment of the assumptions made, their limitations, and their compatibility, which is crucial for assessing the developed model as a whole.
3. Simulation experiment specifications, which we see as a product of the simulation experimentation process, that is crucial for reproducing the experimental results. A simulation experiment specification is everything that allows to precisely repeat the steps that generate the result of the experimentation.

The integration of the Open Provenance Model will provide additional structure to the ODD protocol, and, reaching beyond ODD, it integrates simulation experiments used for calibration and validation.

While the Details block of ODD, especially the *Submodels* element, is concerned with documenting the artifacts, i.e., the submodels and data sources, it does not provide a way to present this information in a structured way. Therefore, a provenance model can complement ODD in two ways, corresponding with the two central goals of ODD – rigor and accessibility. Firstly, it provides the modeler with a guide for documenting these artifacts, providing a framework that contains all artifacts produced. Furthermore, all processes conducted during the development of the simulation model become explicit in the provenance model. These development processes, which are not directly considered by the ODD standard, are equally important to document. Secondly, it can provide the reader with an overview about the various artifacts, including submodels, and data sources, and their relation, making the Details block of ODD more accessible. The explicit documentation of the processes enables the reader to recreate them, which is necessary for exactly reproducing the model.

Simulation experiments, a crucial step in the model development process, form an important part of a simulation model's provenance. For assessing a model's validity it is necessary to know how the model was calibrated and validated, and what data was used in each of these steps. The provenance model will allow retracing crucial steps in the model generation process, and, due to the explicit representation of simulation experiment specifications, facilitate the replication of the experiments. When the experiment specification is even directly executable, the provenance model can be used to build a package containing all models and data needed to replicate the experiment, by packaging the experiment specification with all artifacts the experimentation process depends on.

5 CASE STUDY: A MODEL OF THE DECISION TO MIGRATE

We conduct our case study with a demographic model concerned with the process of forming a decision to migrate. The model (Klabunde et al. 2017b) explores the hypothesis that in a critical phase approximately between the ages of 18 and 40, individuals make a series of important life decision, e.g. to get married or to have children, with which the decision to migrate competes. In the simulation, it is tested whether based on this hypothesis the observed age pattern of migrants can be explained. Thereby, the linked life courses of individuals are in the focus. This includes marriage, fertility, and mortality of individuals, which are governed by stochastic rates, as well as income and expenses. The migration decision process itself is modeled based on the Theory of Planned Behavior (Ajzen 1991). The assumption is, that the decision to migrate is made in multiple stages, through which every potential migrant goes: an intention is formed, plans and then preparations are made, and finally the migration is attempted. Each agent has an intention to migrate, which, in accordance with the Theory of Planned Behavior, is derived from their attitude towards migration, their beliefs about social norms regarding migration, and their beliefs about behavioral control regarding migration. Those three factors are influenced by the agent's personal situation and his or her environment. A total of six free weighting parameters determines the strength with which different aspects influence the migration intention. Finally, the migration intention governs how fast the agent proceeds through the stages of the decision process.

The model was applied to the case of migration from Senegal to Europe. To this end marriage, fertility, mortality, income, and expenses submodels were estimated from data. For marriage a Coale-McNeill model (Coale and McNeil 1972) was fitted, using data from the Demographic and Health Survey of Senegal (DHS) for individuals in Senegal, and from the MAFE survey (Migration between Africa and Europe; Beauchemin 2015) for individuals who migrated. The individuals are then paired by employing a marriage market (Zinn 2012). Fertility was also estimated from DHS and MAFE data. For mortality a Heligman-Pollard model (Heligman and Pollard 1980) was fitted to data from the UN World Population Prospects 2015. Income is taken from IMF data, consumption from World Development Indicators. An initial population was sampled from the 1988 Senegal census. Initial wealth was estimated from data by Davies et al. (2011).

By adjusting the 6 free parameters the model was then calibrated to reproduce the distribution of the age at migration and the proportion of women among the migrants observed in the MAFE survey. Using

the calibrated model they performed further experiments with different scenarios, adjusting the input data for income and fertility.

As can be seen, the migration decision model is made of several components, each derived from an established model from literature fitted to the case of the Senegal using various sources of data. The relation of these components and the data are crucial for the model as a whole. A thorough documentation of the component models, the data sources and their interaction is therefore essential and, due to the complexity of the model, non-trivial.

6 PROVENANCE OF THE SIMULATION MODEL

To demonstrate our approach, we reconstructed the provenance information about the migration decision process model from the publication about the model (Klabunde et al. 2017b), the ODD description (Klabunde et al. 2015), and the information provided together with the model in the OpenABM model repository (Klabunde et al. 2017a).

The result (show in Figure 1), is by no means complete, due to the limited amount of information we have. The migration decision process model itself (*mig. mod.*) is shown as an artifact on the left side of the figure. It was produced through composing its submodels (*comp. model*).

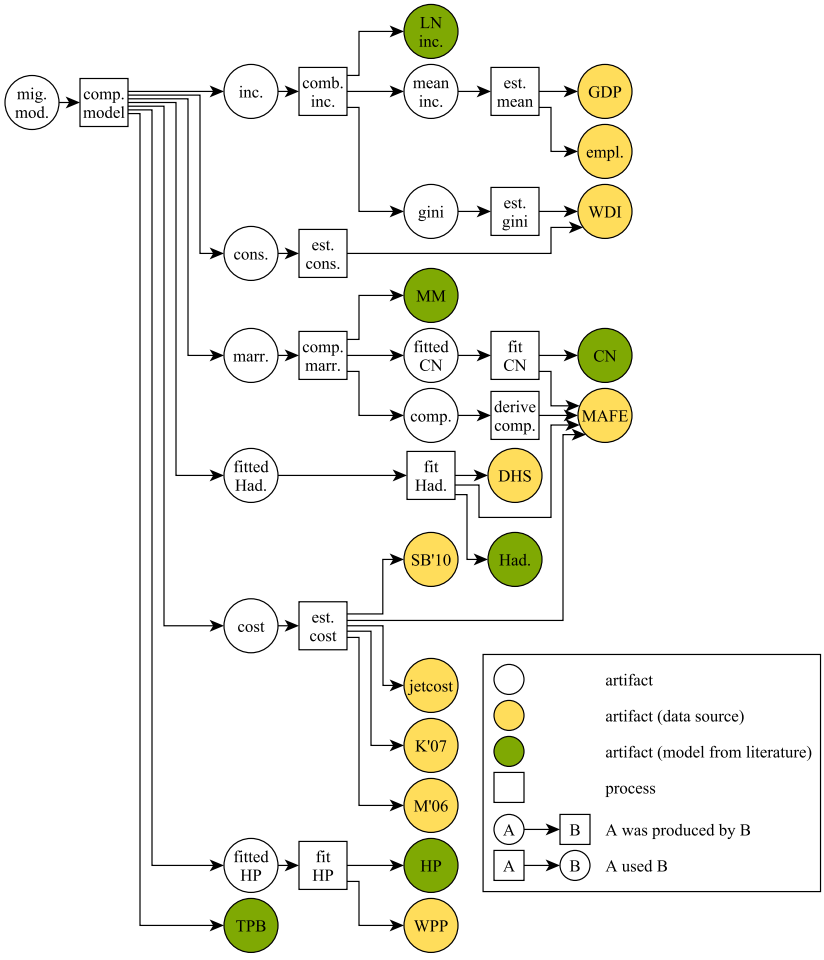

Figure 1: Open Provenance Model for the migration decision model, derived from publications about the model. For a short description of the artifacts and processes in this figure, see Table 2 in Appendix A.

We will now look at one of the components, the mortality model, in detail. The mortaliy model (*fitted HP*) is directly derived from the Heligman-Pollard model of mortality (*HP*), an established statistical mode (Booth and Tickle 2008). It was produced through the process of statistically fitting the Heligman-Pollard model (*fit HP*) to the UN World Population Prospects 2015 data on mortality in Senegal (*WPP*).

For the reader of the ODD document, Figure 1 provides an overview about the model and its components, and put the components into a relation with the literature models and data sources they are derived from. For example, in the provenance model it is made explicit, that the mortality component of the model is derived from the Heligman-Pollard model. That model is widely applied and its validity for different applications has been assessed (Booth and Tickle 2008).

In addition, during model development the structured approach of the Open Provenance Model can help guiding the documentation of the model, especially the Details block of ODD. In the ODD specification of the migration decision model (Klabunde et al. 2015), for example, the mortality component is missing, apart from a remark that mortality is only age- and sex-dependent.

The provenance model does not only consider the artifacts, but also the processes through which they were derived. However, even in the extended documentation, where the mortality component it documented (Klabunde et al. 2017a), it is not clear how the mortality component was produced, i.e., what data was selected from the source, and what methods were employed to fit the model. Similarly, the other fitting processes are not documented in sufficient detail to assess the quality and validity of the simulation model. Therefore, more details about the processes, e.g., in terms of simulation experiments, are needed.

7 SIMULATION EXPERIMENTS AS PART OF A MODEL'S PROVENANCE

Simulation experiments are crucial in developing simulation models. Therefore, they constitute essential information on how a simulation model has been generated. Thus, we have also reconstructed the provenance information for the simulation experiments conducted by Klabunde et al. (Figure 2), including the calibration of the model and the execution of predictive experiments with the calibrated model. The calibration of the model was a two-step process, corresponding to two calibration targets: the proportion of female to male migrants (*sex prop.*), and the age distribution at time of migration (*age dist.*). Both of these targets were derived from the MAFE data set. All of the experiments rely on a suitable initial population (*init. pop.*) which was produced in a multi-step process relying on several data sources (see Figure 2 for details). While the description of the initial state is viewed as a part of the model documentation by the authors of ODD, and therefore included in ODD, we included it in the simulation experiments. This corresponds with the fact that the initial population is experiment-specific, as one might want to use different initial populations for different experiments. Note that the Figure references the artifacts *mig. model*, which is produced in Figure 1 and thus links the two parts of the provenance model, and *MAFE*, which is also identical to the artifact of the same name mentioned in Figure 1.

In the first step of calibration, Klabunde et. al. experimented with the model (*exp. cand.*) to find a set of candidate parameter combinations which can reproduce the proportion of female to male migrants (*sex prop.*) sufficiently well. Apart from the target proportion this process uses the (as yet uncalibrated) migration model (*mig. mod.*) as well as the initial population. The product of this process is twofold: Firstly, of course, the experimentation produces the set of candidate parameter combinations (*cand. set*). The second product is a specification of the experiment, which contains all details needed to reproduce the experimental results. A similar pattern can be seen for the second step of calibration (*exp. age*) and the predictive experiment (*exp. scen.*).

For using the provenance model to ensure reproducibility of the conducted simulation experiments, having the experiment specification as an explicit artifact is crucial. To demonstrate how this can be achieved, we have implemented a calibration experiment similar to *age exp.* in SESSL (Simulation Experiment Specification via a Scala Layer; Ewald and Uhrmacher 2014), see Figure 3. SESSL is an internal domain specific language, based on the general programming language Scala, for the specification of simulation experiments. All experiment specifications in SESSL are valid Scala code and directly

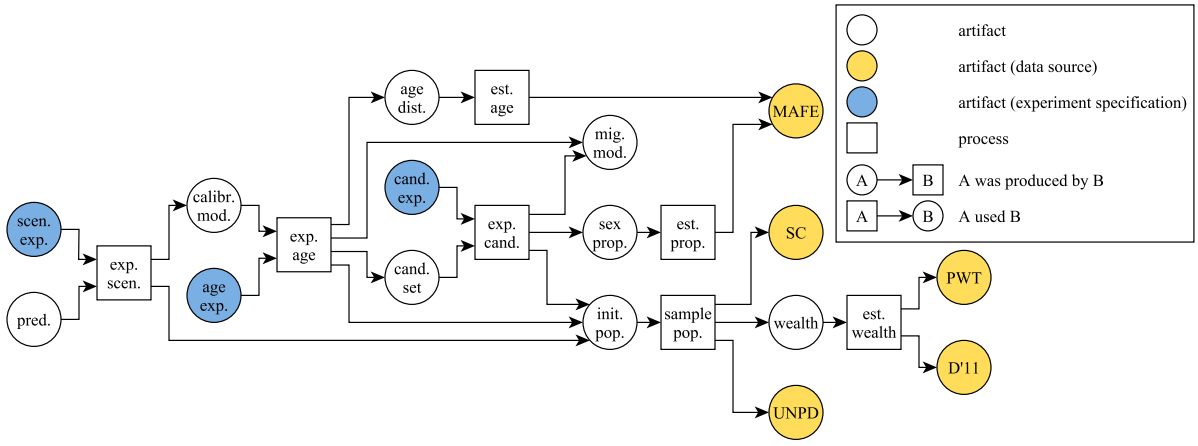

Figure 2: Open Provenance Model for the simulation experiments. Note that the artifacts *mig. mod.* and *MAFE* are identical to the artifacts of the same name in Figure 1. For a short description of the artifacts and processes in this figure, see Table 3 in Appendix A.

executable, allowing for the reproduction of any experiment given its SESSL-specification and the model and data files referenced by it. Using the provenance model and the executable experiment specification we can now build a package that contains all information and data needed to reproduce the experiment. We have to include the experiment specification and all the artifacts used by the experimentation process (similar to Murta et al. 2014).

Provenance information about the simulation experiment can be used, similar to provenance of the model, to trace the origin of data and methods used in the experiment. The origin of the data is especially of interest for validation and calibration experiments, because if the same data used for the calibration of the simulation is also used for validation of the simulation model, the validation result would be invalid. Further, the provenance information allows identifying all artifacts that have been used for executing the simulation experiment. These can be bundled into a container which contains all the data sources, the simulation model and the simulation experiment specification. The container can be then shared to replicate the simulation results.

8 CONCLUSION

With ODD+P we demonstrated how provenance information about simulation models based on the Open Provenance Model (OPM) can form a valuable addition to the ODD protocol. Firstly, the approach guides the modeler in documenting artifacts and processes that contributed to the development of a simulation model. Secondly, it structures the ODD Details block in terms of artifacts, including submodels, and data sources, processes, their relation and roles, adding to the rigor and accessibility of ODD. In our case study on an agent-based migration model, the provenance information supported an efficient communication of the mortality component, and the simulation model as a whole. It enabled us to reconstruct assumptions made by the submodels, the theories they were derived from, methods used for developing them, and data sources used for fitting parameters of the simulation model.

With a standardized provenance model such as OPM, inference mechanisms such as the OPM-level query language OPQL (Lim et al. 2011), can be employed to answer questions about the provenance of a simulation model. For example if we identify an methodological error in the collection of a dataset we can use the provenance model to infer all affected model components which need to be revised.

We have derived the provenance information post factum from model documentation. Ideally the provenance information is collected directly during a simulation study. Different methods such as version

```
1  minimize{ (params, objective) => execute(
2    new Experiment with ParallelExecution with ParameterMaps with Observation {
3      model = "migration/migration.ml3"
4      simulator = NextReactionMethod()
5      parallelThreads = -1
6      replications = 1
7      initializeWith(new JsonStateBuilder("migration/initialstate2000.json"))
8      startTime = 1982
9      stopTime = 2050
10
11     fromFile("migration/maleMortality.csv")(); fromFile("migration/femaleMortality.csv")()
12     fromFile("migration/fertility.csv")(); fromFile("migration/income.csv")()
13     fromFile("migration/ageDifferenceModifier.csv")(); fromFile("migration/baseMarriageRate.csv")()
14     fromFile("migration/borderEnforcement.csv")(); fromFile("migration/disc.csv")()
15
16     set("minFertilityAge" <~ 12, "maxFertilityAge" <~ 49, "ageOfAdulthood" <~ 16, "ageOfRetirement" <~ 65,
17         "minMarriageAge" <~ 9, "maxMarriageAge" <~ 60, "meanMigrationStartAge" <~ 17,
18         "spouseAgeModifier" <~ -0.01301431, "intercept" <~ -0.490129556,
19         "homeCountryGini" <~ 0.4, "hostCountryGini" <~ 0.3)
20
21     for ((param, value) <- params.values) set(param <~ value)
22
23     observeAt(Change(agentType = "Person", field = "migrationStage",
24       filter = "ego.migrationStage = 'migrated' && ego.planningTime != 0")) {
25       observe("migrationAge" ~ expression("ego.age"))
26     }
27
28     withReplicationsResult { replication =>
29       val ages = values[Double](replication, "migrationAge")
30       val relFreqs = ages.groupBy(age => age.toInt).mapValues(_.size.toDouble / ages.size)
31       objective <~ util.math.Misc.mse(
32         CalibrationData.referenceAges,
33         Range.inclusive(0, 52).map(relFreqs.withDefaultValue(0.0)))
34     }
35
36     def values[T](result: ObservationReplicationsResultsAspect, name: String): Iterable[T] =
37       for (run <- result.runs if run ? name; value <- run.values[T](name)) yield value
38   })
39 } using new Opt4JSetup {
40   param("xi", 0.05, 0.1, 0.5)
41   param("zeta", 1.0, 10.0, 100.0)
42   param("alpha", 0.0005, 0.0005, 0.002)
43   param("beta", 50.0, 100.0, 1000.0)
44   param("gamma", 0.0001, 0.0005, 0.001)
45   param("rho", 0.05, 0.1, 0.02)
46   optimizer = ParticleSwarmOptimization(particles = 1, iterations = 1)
47
48   withOptimizationResults(results => println("Overall results: " + results.head))
49 }
```

Figure 3: Specification of a calibration experiment in SESSL. In line 1 it is stated, that in the experiment an objective shall be minimized. The objective is the result of the simulation experiment specified in line 2–38. In line 2 a Scala object that represents a simulation experiment is created. The experiment uses parallel execution and parameter maps, which represent time-series parameters, and contains observations. In line 3–9 the model file is set and the simulator configured. In line 11–19 the fixed model parameters are set. This includes time-series parameters, e.g., for age-dependent mortality rates, as well as simple parameter constants. Line 21 sets the varied model parameters to the values the optimization algorithm demands. In line 10–16 the objective of optimization, the mean squared error between the simulated age distribution and the age distribution derived from the MAFE data, is calculated. The latter is provided by the Scala object `CalibrationData`. In line 18–19 a function is defined, that is used in the calculation of the objective.

control systems, scientific workflows, experiment scripts and domain specific languages allow for either intrusively, with further action from the modeler required, or non-intrusively, collecting crucial provenance information of simulation models (Ruscheinski and Uhrmacher 2017; Murta et al. 2014). However, they have still to be brought together effectively to support establishing provenance as salient meta-data of simulation models, and thus increasing trust into the quality and validity of simulation models.

ACKNOWLEDGMENTS

This research was supported by the German Research Foundation (DFG) via research grants UH-66/15 and UH-66/18. The authors thank Tom Warnke for his assistance with SESSL experiments.

A APPENDIX

Table 2: Artifacts and processes shown in Figure 1.

Id.	Description	Reference
CN	Coale-McNeil model of transition rates to marriage	(Coale and McNeil 1972)
comb. inc.	combination of the estimated income measures to derive estimated income distributions for Senegal and France	(Klabunde et al. 2017a)
comp.	partner compatibility measure to use by the marriage market	(Klabunde et al. 2015)
comp. marr.	composition of the different components of the marriage model	(Klabunde et al. 2017a; Klabunde et al. 2017b)
comp. mod.	composition of the different components of the migration decision model	(Klabunde et al. 2017b)
cons.	consumption time series for Senegal and France	(Klabunde et al. 2017a)
cost	migration cost model	(Klabunde et al. 2017a)
derive comp.	derivation of a compatibility measure from the MAFE data	(Klabunde et al. 2015)
DHS	Senegal Demographic and Health Survey 1986 - 2014	
empl.	IMF employment data	
est cons.	estimation of a consumption time series	(Klabunde et al. 2017a)
est. cost	estimation of the mean migration cost as a weighted average of the migration cost when using different modes of transit	(Klabunde et al. 2017a)
est. gini	estimation of Gini indices in Senegal and France	(Klabunde et al. 2017a)
est. mean	estimation of a mean income time series in Senegal and France	(Klabunde et al. 2017a)
fit CN	fitting of the Coale-McNeil model to the Senegal data	(Klabunde et al. 2017a; Klabunde et al. 2017b)
fit Had.	fitting of the Hadwiger model to the Senegal data	(Klabunde et al. 2017a)
fit HP	fitting of the Heligman-Pollard model to the Senegal data	(Klabunde et al. 2017a)
fitted CN	fitted Coale-McNeil model	(Klabunde et al. 2017a)
fitted Had.	fitted Hadwiger model	(Klabunde et al. 2017a)
fitted HP	fitted Heligman-Pollard model	(Klabunde et al. 2017a)
GDP	IMF GDP data	
gini	Gini index estimation for Senegal and France	(Klabunde et al. 2017a)
Had.	Hadwiger model of fertility rates	(Hadwiger 1940)
HP	Heligman-Pollard model of age-dependent force of mortality	(Heligman and Pollard 1980)

Table 2: (continued)

Id.	Description	Reference
inc.	lognormal income distributions for Senegal and France at multiple times	(Klabunde et al. 2017a)
jetcost	flight cost data retrieved from jetcost.de (specifics unknown)	
K'07	estimations of the cost of migration by boat from Senegal to the Canary Islands by Kohnert	(Kohnert 2007)
LN inc.	lognormal distribution as model for income distributions	(Bandourian et al. 2002)
M'06	estimations of the cost of the migration by boat from Senegal to various points in Europe by van Moppes	(van Moppes 2006)
MAFE	Migration from Africa to Europe dataset (see also Table 3)	(Beauchemin 2015)
marr.	marriage model component	(Klabunde et al. 2017a)
mean inc.	mean income time series for Senegal and France	(Klabunde et al. 2017a)
mig. mod.	migration decision process model (see also Table 3)	(Klabunde et al. 2017b)
MM	marriage market for matchmaking	(Zinn 2012)
SB'10	estimation of the cost for smugglers for illegal migration by plane and ship by Schmid and Borchers	(Schmid and Borchers 2010)
TPB	Theory of Planned Behavior as model of a decision process	(Ajzen 1991)
WDI	World Development Indicators	
WPP	UN World Population Prospects 2015	

Table 3: Artifacts and processes shown in Figure 2.

Id.	Description	Reference
age dist.	age distribution of migrants	(Klabunde et al. 2017b)
age exp.	calibration experiment for calibrating the age distribution of migrants produced by the model	(Klabunde et al. 2017b)
calibr. mod.	calibrated migration decision model	(Klabunde et al. 2017b)
cand. exp.	candidate selection experiments, selecting parameter combinations which reproduce the proportion of female migrants	(Klabunde et al. 2017b)
cand. set.	set of candidate parameter combinations	(Klabunde et al. 2017b)
D'11	Davies et al. about the level and distribution of global household wealth	(Davies et al. 2011)
est. age	estimation of the age distribution of migrants	(Klabunde et al. 2017b)
est. prop.	estimation of the proportion of female migrants	(Klabunde et al. 2017b)
est. wealth.	estimation of the wealth distribution in the initial population	(Klabunde et al. 2017a)
exp. age	age calibration experimentation	(Klabunde et al. 2017b)
exp cand.	candidate selection experimentation	
exp. scen.	scenario experimentation with modified income growth in Senegal	(Klabunde et al. 2017b)
init pop.	initial population for the simulation	(Klabunde et al. 2017a)
MAFE	Migration from Africa to Europe dataset (see also Table 2)	(Beauchemin 2015)
mig. mod.	migration decision process model (see also Table 2)	(Klabunde et al. 2017b)

Table 3: (continued)

Id.	Description	Reference
pred.	predicted rate of migration with modified income growth in senegal	(Klabunde et al. 2017b)
PWT	Penn World Table 6.1	
sample pop.	sampling of an initial population from the census data	(Klabunde et al. 2017a)
SC	Senegal Census 1988	
scen. exp.	experiment with modified income growth in senegal	(Klabunde et al. 2017b)
sex prop.	proportion of female migrants	(Klabunde et al. 2017b)
UNPD	UN Population Division data about the age structure of Senegalese in France in 1982	
wealth	estimated distribution of household wealth in Senegal	(Klabunde et al. 2017a)

REFERENCES

Ajzen, I. 1991. "The Theory of Planned Behavior". *Organizational Behavior and Human Decision Processes* 50(2):179–211.

Bandourian, R., J. McDonald, and R. S. Turley. 2002. "A Comparison of Parametric Models of Income Distribution Across Countries and Over Time". SSRN Scholarly Paper ID 324900, Social Science Research Network, Rochester, NY.

Beauchemin, C. 2015. "Migration between Africa and Europe (MAFE): Looking beyond Immigration to Understand International Migration". *Population* 70(1):13–38.

Booth, H., and L. Tickle. 2008. "Mortality Modelling and Forecasting: A Review of Methods". *Annals of Actuarial Science* 3(1-2):3–43.

Callahan, S. P., J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. 2006. "VisTrails: Visualization Meets Data Management". In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, edited by S. Chaudhuri et al., 745–747. New York City, New York: ACM.

Coale, A. J., and D. R. McNeil. 1972. "The Distribution by Age of the Frequency of First Marriage in a Female Cohort". *Journal of the American Statistical Association* 67(340):743–749.

CoMSES Network 2018. "CoMSES Computational Model Library (OpenABM)". https://www.comses.net/. Accessed March 23rd, 2018.

Davies, J. B., S. Sandström, A. B. Shorrocks, and E. N. Wolff. 2011. "The Level and Distribution of Global Household Wealth". *The Economic Journal* 121(551):223–254.

Ewald, R., and A. M. Uhrmacher. 2014. "SESSL: A Domain-Specific Language for Simulation Experiments". *ACM Transactions on Modeling and Computer Simulation* 24(2):11:1–11:25.

Grimm, V., U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse, A. Huth, J. U. Jepsen et al. 2006. "A Standard Protocol for Describing Individual-Based and Agent-Based Models". *Ecological Modelling* 198(1):115–126.

Grimm, V., U. Berger, D. L. DeAngelis, J. G. Polhill, J. Giske, and S. F. Railsback. 2010. "The ODD protocol: A review and first update". *Ecological Modelling* 221(23):2760–2768.

Groth, P., and L. Moreau. 2013. "PROV-Overview – An Overview of the PROV Family of Documents". https://www.w3.org/TR/prov-overview/, World Wide Web Consortium. Accessed March 23rd, 2018.

Hadwiger, H. 1940. "Eine Analytische Reproduktionsfunktion Für Biologische Gesamtheiten". *Scandinavian Actuarial Journal* 1940(3-4):101–113.

Heligman, L., and J. H. Pollard. 1980. "The Age Pattern of Mortality". *Journal of the Institute of Actuaries* 107(1):49–80.

Klabunde, A., F. Willekens, S. Zinn, and M. Leuchter. 2015. "An Agent-Based Decision Model of Migration Embedded in the Life Course – Model Description in ODD+D Format". MPIDR working paper WP-2015-002, Max Planck Institute for Demographic Research, Rostock, Germany.

Klabunde, A., S. Zinn, F. Willekens, and M. Leuchter. 2017a. "Multistate Modeling Extended by Behavioral Rules (Version 1.5.0)". https://www.comses.net/codebases/5146/releases/1.5.0/, CoMSES Computational Model Library. Accessed July 30th, 2018.

Klabunde, A., S. Zinn, F. Willekens, and M. Leuchter. 2017b. "Multistate Modelling Extended by Behavioural Rules: An Application to Migration". *Population Studies* 71(sup1):51–67.

Kohnert, D. 2007. "African Migration to Europe: Obscured Responsibilities and Common Misconceptions". GIGA Working Paper 49, German Institute of Global and Area Studies, Hamburg, Germany.

Lim, C., S. Lu, A. Chebotko, and F. Fotouhi. 2011. "OPQL: A First OPM-Level Query Language for Scientific Workflow Provenance". In *Proceedings of the 2011 IEEE International Conference on Services Computing*, edited by H. A. Jacobsen et al., 136–143. Piscataway, New Jersey: IEEE.

Moreau, L., B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. V. den Bussche. 2011. "The Open Provenance Model Core Specification (v1.1)". *Future Generation Computer Systems* 27(6):743–756.

Murta, L., V. Braganholo, F. Chirigati, D. Koop, and J. Freire. 2014. "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, Switzerland: Springer.

Rahmandad, H., and J. D. Sterman. 2012. "Reporting Guidelines for Simulation-Based Research in Social Sciences". *System Dynamics Review* 28(4):396–411.

Ruscheinski, A., and A. Uhrmacher. 2017. "Provenance in Modeling and Simulation Studies – Bridging Gaps". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan et al., 872–883. Piscataway, New Jersey: IEEE.

Schmid, S., and K. Borchers. 2010. "Vor den Toren Europas? Das Potenzial der Migration aus Afrika". Report No. 7, Federal Office for Migration and Refugees, Nürnberg, Germany.

Schulze, J., B. Müller, J. Groeneveld, and V. Grimm. 2017. "Agent-Based Modelling of Social-Ecological Systems: Achievements, Challenges, and a Way Forward". *Journal of Artificial Societies and Social Simulation* 20(2):8.

Simmhan, Y. L., B. Plale, and D. Gannon. 2005, September. "A Survey of Data Provenance in e-Science". *ACM SIGMOD Record* 34(3):31–36.

van Moppes, D. 2006. "The African Migration Movement: Routes to Europe". Report No. 5, Working Papers Migration and Development, Radboud University, Nijmegen, Netherlands.

Wolstencroft, K., O. Krebs, J. L. Snoep, N. J. Stanford, F. Bacall, M. Golebiewski, R. Kuzyakiv, Q. Nguyen, S. Owen, S. Soiland-Reyes, J. Straszewski, D. D. van Niekerk, A. R. Williams, L. Malmström, B. Rinn, W. Müller, and C. Goble. 2017. "FAIRDOMHub: A Repository and Collaboration Environment for Sharing Systems Biology Research". *Nucleic Acids Research* 45(D1):D404–D407.

Zinn, S. 2012. "A Mate-Matching Algorithm for Continuous-Time Microsimulation Models". *International Journal of Microsimulation* 5(1):31–51.

AUTHOR BIOGRAPHIES

OLIVER REINHARDT is a Ph.D. student in the modeling and simulation group at the University of Rostock. His email address is oliver.reinhardt@uni-rostock.de.

ANDREAS RUSCHEINSKI is a Ph.D. student in the modeling and simulation group at the University of Rostock. His email address is andreas.ruscheinski@uni-rostock.de.

ADELINDE M. UHRMACHER is a professor at the Institute of Computer Science, University of Rostock and head of the modeling and simulation group. Her email address is adelinde.uhrmacher@uni-rostock.de.