Agent Based Modelling for Social Research

Provenance for Simulation Studies

Hello, everybody now to short course on provenance for simulation study, you all heard from Andrei about the ODD protocol and how you are reporting your results and are reporting what you did, more or less in this kind of documentation standard. And of course, it provides a certain kind of structure, however, you do not get a really easy overview about the development process and also, it is not represented in a computer variable or accessible format.

So at the end of the day, if you doing this, so you have a nice simulation model developed as, for example, our cooperation partners from the Max Planck Institute of Rostock, and then you writing a paper and then you're then starting to document this model, with the hub of like the ODD report, and also reports referring to the inputs that you use, at the end of the day, you have a lot of papers that you need to sift through. So to get to the information you are interested. And the question is, of course, can we do better.

And simulation studies are of course, really intricate processes similarly like to maybe not as complicated migration process, but nonetheless, you have a lot of data that you have to collect, you have hypothesis, you have series, and you have to put everything together into your model. And this kind of process, of course, can also be subject to modelling so to better kind of understand what you did. And hear what you did or how our simulation model came into being. This information we typically called provenance or as it is stated here. So, it is 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. So, that is what we would like to have, because of course, we want to have trust in our models and our simulation models, and we want to do so, as I said, in a structured computer accessible manner.

So, what we then use here is profit profit is a standard to describe provenance. Actually, it has not been developed to help us deal with our simulation studies, but it is a rather broad area of research in databases and reproducible science. So, PROV is a standard and to describe provenance information as an annotated causality crafts we have a craft which nodes and edges and we have entities as the nodes are the products, the inputs and our simulation studies and activities. So actions performed on or caused by entities resulting in new entities and to see clearly what connects the relations between entities activities are that an activity uses an entity, it may use it in different kinds of roles, like for example, our simulation stuff, as in kind of like when you are doing the, you're using data for calibration, you are using data for validation. And we have entities that are generated by these activity for example, that you have now received or that you achieved a validated simulation model.

So, the idea back to us is now to apply this kind of young, more structured or more computer accessible way of describing what we did with our migration model. So, we are back in our migration study, not ours, but my colleagues at the Max Planck Institute and we went through the ODD definition that they did and more or less took these things apart. And what you see here, as I said is you have an activity and created by this activity is the migration model. So, this on the right side is more or less time flows from from the left to the right. And what you see here is at the, at the right side, the result is the migration model. The activity is you composing the model from different kinds of processes. And you find here a lot of different processes that I wanted to go into detail here. But what do you see is that income is obviously approach, something which is modelled here explicitly. Also the what you're consuming as a result of your daily life then had whether you are marrying as a marriage processes, fertility processes, all these processes you typically expect in these kind of models, then there exists also costs. And these costs here are the costs that are induced you to your wish to migrate, and you have of course, a certain kind of mortality. And then the decisions that, again has to kind of like this decision model that all come together to form now, the migration model.

This kind of like, here, the decision here at that point, one moment, the decision here is, of course, depending on the theory of planned behaviour, and takes, of course, into account everything what else is currently the state of our agent. But nevertheless, let's look a little bit more in more concretely how these things are put together or how we, how we develop this kind of of provenance model of what has happened here.

So, we are starting with the birth So, this is a fertility. Young, were in the ODD document, there is a lot of description how these fertility model is supposed what it should cover. Then, when we are looking more closely, it's, it's dependent on the fit Had model for each dependant fertility, some of you might know that, and of course, on certain kind of data, and this kind of this here is thse data of senegal and these are now then what you're doing is that you are now fitting this kind of the the head bigger model to depending on these kind of data. So, to achieve a fertility model, but you want also to know a little bit about the fertility of the people living in France and for that, you are as people migrated to France and therefore, you use a different kind of data data set that is based on interviews, the main data set, and again you are fitting it, so, you fertility model at the end of the day consists of two things, then the fertility is determined depending on whether you are currently in the senegal or you are in France, based on the on the data there and together then on the basis of the the head figure model, okay, but as I just said, This fitting, the fitting is described rathrt extensively here.

And fitting of course, is a kind of simulation experiment that you are conducting with your simulation model. And to have a little bit more information about that actually, that they used are kind of, certain kind of fitting procedure that you are getting out of this craft. So we are getting into that. So we have now when you're looking here, you have here now a migration model and the first fitting is referring to the gender and out of this falsls specification of the simulation experiment. So this, this here is a specification of the simulation model, the simulation experiment that you can rerun, depending on how you encoded it maybe as a Python code, maybe as an R script, or maybe as as a specific language, like, for example, example, sesa and you have then a certain kind of model, and again, this model is subject to further fitting to an age distribution down here, and you're getting out of this a calibrated model and the age experiment here, and then at the end of the day, you are doing certain kinds of experiments here the scenario experiment. And those will then produce the predictions that are then shown in the figures of in your paper. So that is, more or less how you can see how these, these provenance crafts gives you detailed information about what you did. And the good thing about that. If you have this as a provenance craft, of course, then you can query it. Yeah.

But first of all, how does this provenance craft, is different from the typical provenance information as said, we are using a standard for provenance which has not been defined for simulation studies, but we are adapting it simply by saying okay what do you need if you want to describe a simulation study. As entities you have possibly requirements assumptions, quantitative models, a simulation model, simulation experiment data or theory, you can think of other things as well but maybe that is a good point to start.

The activities that you have a building simulation model possibly calibrating, validating simulation model analysing it. And you can have also statistical fitting if you want to have also look a little bit how you how you're working with the data, the roads are referring when you're using in your activities, you're using data for example, for calibration validation, you're using simulation models, possibly also for cross validation. So, you can think about these things.

And if you have now this kind of of information, then you can put this in a database, preferable craft database and then you can for example, ask things like okay, all the data are invalidated which product of my simulation study needs to be revisited. So, where, where could I possibly assume that a data set becoming invalid, what kind of products of my simulation study become invalid as well? and you see here that there are certain kinds of data that I derived here and this certain kind of processes, yeah, certain kinds of models that I did. You see, these are the experiments the simulation experiments is defined, all of those have to be checked again.

You can because of course, these these crafts become tend to become rather large and maybe, to have a better overview can of course, do are these kind of that you're thinking about hierarchies or that you have a modular approach. So that for example, here, the activity A4 is an activity that again is composed of the different kinds of activities using different kind of real data here with W3 or W indicated and certain kinds of models and then you are then again here for example, A4, A4 is again a composed activity, so you can zoom in as you then would like to do, okay.

So, and when you have then very, very large provenance classes, this is a small one that is really a toy example, but it has already more than 70 nodes and 119 edges what you then can do is of course, query all input data used in the last model version, for example, and then you're getting the different kind of here parameter values out of that and then where they are coming from, or building and validating phases not only to show those, yeah, or birth as few on the final results of the simulation study only the last versions of entities are shown and you see here that it was really very simple simulation study it was a toy of a predator prey. And we used only two publications here to build first the conceptual model, then come up with a kind of requirement and then this kind of requirement to show that it really, the simulation model had done to to adhere to these kind of behavioural requirements, which you can state for example, in terms of data, so the data should be reproduced. Or for example, that you saying okay, you would like to see a peak of that height at a certain time, things like that. Okay.

What you can also do if you have now provenance, about one simulation model, how you build one simulation model here how you build run simulation model, you can then relate it to other simulation models that you base your, your simulation model on. For example, if you're thinking about here, this kind of, here down here, one moment. one monebt I have to check, this is not easy. So down here sample, this is yeah, this yeah, no. So you're using different kinds of data here. And then what you see here you are using also, other, other, other simulation as you were referring to other simulation studies, yeah, reusing for example models, or, yeah, part of the models are cross validating your simulation model with other models. Because you know what these kind of models have been built based upon, you can assess whether this reuse of simulation model is, for example valid for the questions and the data you have.

And you can even go beyond simulation studies actually that we are we are doing or we did in the in our BAPS project, that Jacob Bijack is the is the head off. So here what we used here in the, here, you see in the beginning, you see the different kind of iteration of the different types of simulation models that we developed. And you see, at the end, it here, this point here, this really uses data that have been cleaned up, and that have been somehow modified during these kind of processes here, which part of the project is responsible for, and they have been enriched by, by our results from psychological experiments, that another part of the project did. So at the end of the day, these kind of approach allows you a kind of, you really to, to keep on top of what you're doing during your project, particularly if you have a project where you are developing simulation models yeah, in different versions, of course of them may be different simulation models, you have to deal with data and her in and this case, also with real experiments.

Of course, some kind of software support is is important, as I said, when you want to do that, you have a database, and preferably one craft based database like Neo4J is very make sense. And this is kind of like an user interface for the web we developed. And that was we I mean, Kai Budda and a student from the University of New Brunswick and Canada. And where you then can add in some of this information, but it's still a research prototype, I have to admit

So PROV Ontology for modelling and simulation studies, what have we gained, it's a structured approach and an accessible approach to conduct simulations that is accessible. First of all, from the point of view that you can more easily see what the structure is and how things have evolved. Here, so it's an essential ingredients, I have to admit is here to making all sorts of simulation experiment explicit that is very important, because the simulation experiments are typically the point where you are getting the simulation model and the data together, for example, for fitting, or also for validating, so to have those explicitly described, and better, not in text, but really in terms of, for example, the script and so forth, of course, you can annotate it, but nevertheless, that should be also part of that. And this kind of information, it grows of course, depending on the complexity of your endeavour, then you can have an accessible for filtering, analysing and querying. And, for example, are the same data used for calibration or validation, these are queries that you would like to ask. And it's really interesting, particularly if you're building simulation families as a simulation, model families, for example, about migrations, yeah. So then you are referring to other migration service. And then you can think about this and that you have really a network of information.

Yeah, so the usefulness so far of our approach has been demonstrated in several simulation studies, cell biology, ecology, and demography. And we have to say it really kind of like all modelling efforts, it really structures, the knowledge and it really structured the process in which you are conducting a simulation study. However, what we should say is, this is really a research prototype. So as a software, you're welcome to try it out a little bit. But it's nothing that we can really support you because we are at university, so, but nevertheless, so there is a little bit of support but as still I would really like to encourage you to try it out. Because it's really it helps a lot to get some structure in what you did, what kind of data you use for what kind of purposes but it is still ongoing research. And so whereas ODD isn't established standard. At the end of the day, of course, you want to bring everything together and for now, I'm looking forward to your questions and thank you for the attention.