Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

Towards metadata standards for sharing simulation outputs

Gary Polhill
The James Hutton Institute,
Craigiebuckler, Aberdeen.
AB34 5LU, UK
Email: gary.polhill@hutton.ac.uk

Terry Dawson
Centre for Environmental Change and Human Resilience,
University of Dundee,
Perth Rd., Dundee. DD1 4HN, UK
Email: t.p.dawson@dundee.ac.uk

Dawn Parker, Xiongbing Jin, Kirsten Robinson
University of Waterloo,
200 University Avenue West,
Waterloo, ON N2L 3G1, Canada
Email: {dcparker, x37jin, k4robins}@uwaterloo.ca

Tatiana Filatova, Alexey Voinov
University of Twente,
Enschede, The Netherlands
Email: t.filatova@utwente.nl, aavoinov@gmail.com

Michael Barton
Arizona State University, School of Human Evolution and Social Change, PO Box 872402, Tempe, AZ 85287-2402, USA
Email: michael.barton@asu.edu

Edoardo Pignotti, Peter Edwards
University of Aberdeen,
Aberdeen. AB24 3UE, UK
Email: {p.edwards, e.pignotti}@abdn.ac.uk

***Abstract*—This extended abstract outlines a prototype metadata standard for recording outputs of social simulations, to be refined as part of a project funded through the third round of the Digging into Data challenge. This is with a view to gathering community feedback on the proposals.**

I. Introduction

AGENT-BASED models, perhaps more than other kinds of simulation modelling tool, are capable of producing large quantities of simulation data. This is not just because, through representing individuals and their interactions explicitly, there will be data in each time step of the model for each agent and each interaction, but also because a typical use-pattern for agent-based models is to run them multiple times to test the behaviour of the model under different parameter settings and seeds for pseudo-random number generators.

As a consequence, there are challenges associated with interpreting, analysing and visualising results from agent-based models that are akin to those of the 'big data' community with social datasets collected from the real world. Traditional methods for mapping relationships among input, parameter and output variables, such as regression models, are generally designed for independent variables that are additively linear (even when interaction terms are included), generally assume continuous and monotonic dependent variables, and are developed under the assumption that dependent and independent variables have Gaussian distributions. However, data outputs from simulations of complex systems are often nonlinear with discontinuities, discrete and leptokurtic.

The MIRACLE project (Mining Relationships Among Variables in Large Datasets from Complex Systems) will be investigating and developing tools for analysing and visualising outputs from social simulations. We propose to build tools on the CoMSES Net web platform allowing users to upload their model output and associated metadata, visualise and analyse results, and conduct comparative meta-analyses. As part of this exercise, we wish to engage with the social simulation community to gather requirements.

Model output metadata is information describing outputs from simulations. These outputs could take various forms and be stored in different file formats, including screenshots or videos of the model running, data from time-series graphs, networks or spaces, and detailed data describing the states of individual agents. Model output metadata provides more information about the files than may be captured in the file formats themselves, which could include how the files were generated (version of model implementation software, input parameters), why (e.g. as part of an experiment, or for a presentation or publication), or even simply that two files were generated by the same run of the model.

The main purpose of the MIRACLE project is to make data analysis and visualisation tools available and easy to use to practitioners in agent-based modelling. However, since some of the metadata we might record about simulation outputs pertains to its provenance (i.e. the parameter settings, model versions, etc. that generated it), the project should have the added bonus of facilitating replicability and increasing transparency about how the results were generated for stakeholders in the model [1].

II. Earlier work

Early work on using metadata to record simulation outputs includes [2], which used an OWL ontology [3] to record the outputs of experiments with FEARLUS [4]. The ontology is shown in Fig 1; specific subclasses of the ontology were used to describe FEARLUS models, simulation outputs, experiments and results. The ontology provides the ca-

≥ This work was sponsored by the third round of the Digging into Data challenge. http://www.diggingintodata.org/

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

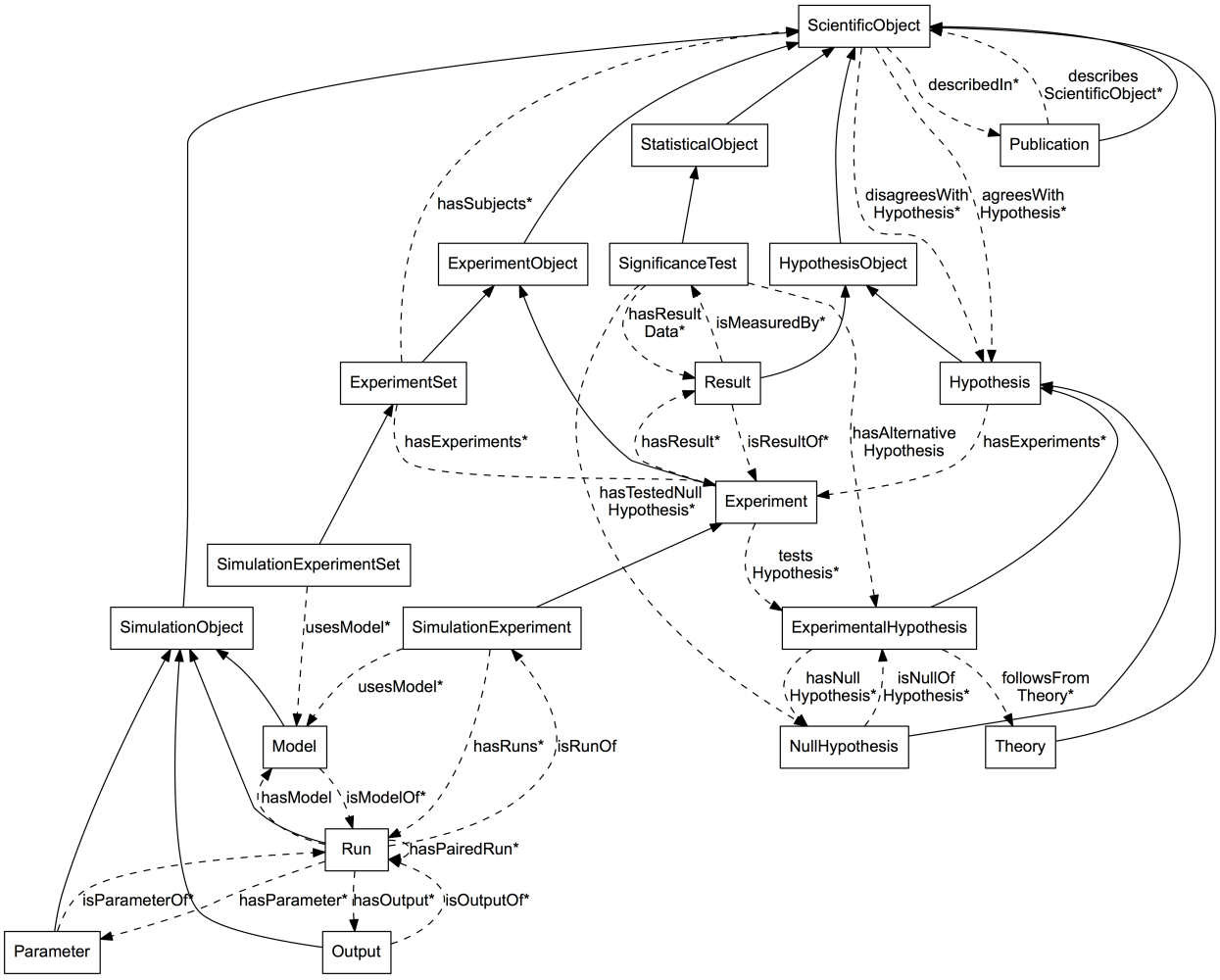

Fig 1: The FEARLUS-G ontology used for recording outputs of simulation experiments. Solid lines connect subclasses to their superclasses; dashed lines are relationships among classes, and are accompanied by a label that ends with an asterisk if the relationship is many-one.

pability to record rich metadata about how and why simulation `Outputs` were generated, and was used as part of a system in which some of the metadata (in particular, instances of `Result`, `Run` and `Output` and relationships among them) were generated automatically.

The ontology was designed with hypothesis-driven research using simulation experiments in mind, and applying conventional statistical tests. This kind of approach has been the subject of some criticism [5], as (i) modifying the input or algorithm of a simulation for the purposes of testing a null hypothesis arguably means the tests are not being conducted on the same 'world'; (ii) there is now sufficient computational power that significant results can be obtained just by performing more runs. Thus, although the ontology could be reused, there are good reasons to consider more general metadata about the *provenance* of simulation output data.

More recent work [6] has explored the use of the PROV ontology [7] as a foundation for recording metadata about how simulation outputs are generated, distinguishing three types of provenance, of which (2) is the most relevant here: (1) the social process of constructing the model; (2) running the model; (3) 'history' within a simulation run. The PROV ontology has entities, activities and agents as the main classes, and according to [6], entities for type (2) provenance are the data and parameters used to initialise or as input to the model (including where they have come from), software used to run it, and output files generated by it; activities are running the model; and agents the hardware and users.

III. Towards a metadata standard

A specialisation of the PROV ontology drawing on the FEARLUS-G ontology seems an appealing starting point for recording metadata about simulation outputs. Our prototype is shown in Fig 2. The FEARLUS-G `hasModel` and `hasParameter` properties are subproperties of the PROV `used` property, and `hasOutput` is a subproperty of `generated`. Properties can also be defined to show the relationship between the `Hardware` and `User`

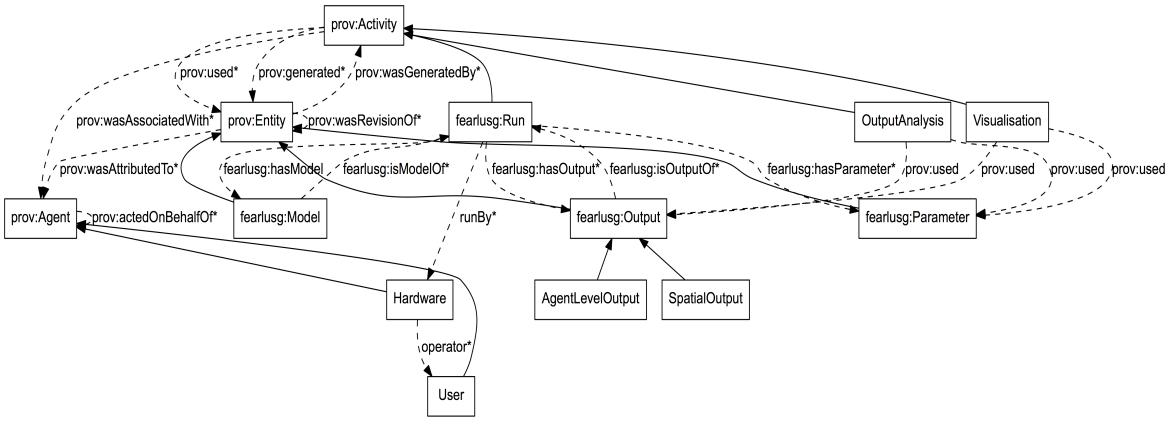

Fig 2: Specialisation of the PROV ontology using FEARLUS-G as a starting point for a simulation output metadata standard. Specialisations of PROV drawing on FEARLUS-G have a fearlusg `namespace`, PROV entities have a `prov` namespace; additional proposed entities have no specified namespace.

`prov:Agents`, with the `operator` property a subproperty of `actedOnBehalfOf`. Similarly, the `runBy` property is a subproperty of `wasAssociatedWith`, used to record the `Hardware` on which the `Run` was conducted.

The PROV ontology provides further vocabulary that might be useful, such as `wasRevisionOf`, which could be used to record versions of `Models`, and `wasAttributedTo`, which could record authorship of `Models`.

In the MIRACLE project, however, there will be a need to record further metadata associated with other possible subclasses of `Activity`, as reflected in the classes `OutputAnalysis` and `Visualisation`. It will also be important to record different kinds of output – in Fig 2 we have shown two for illustration: `SpatialOutput` and `AgentLevelOutput`. Though not shown in the diagram, certain subclasses of `OutputAnalysis` and `Visualisation` are associated with specific subclasses of `Output`. For example, `SpatialOutput` would be suitable for geostatistical subclasses of `OutputAnalysis`. At a workshop at iEMSs 2014 in San Diego, participants identified space-time visualisation tools as a particularly high priority, and clearly such tools would only operate on `SpatialOutput` variables.

An ontology such as the proposed would enable the user to understand whether their output is suitable for analysis or visualisation using a particular technique, and to use standard provenance reasoning techniques to gain some understanding of the processes by which the output was generated.

Workflow tools can prove useful in capturing the processes by which simulation outputs were generated and analysed, and provide a metadata basis for recording scientists' *intent*: goals and constraints of the research [8].

IV. Discussion

A common problem with recording metadata, is that manual entry can be a barrier to its adoption [9]. There is a gap between what would ideally be recorded, and what can feasibly be recorded, either through software facilitation (where constraints might pertain to computational cost or disk space associated with computing and storing relevant metadata) or through manual form entry during upload (where constraints might pertain simply to the amount of metadata users are willing to enter at any one time).

Work with the virtual research environment reported in [9] also noted resistance of users to a static metadata framework, and the authors developed a hybrid semantic/social web approach allowing interoperation between community-driven metadata (e.g. tags) and formal assertions in OWL that are amendable to automated reasoning. It is reasonable to anticipate that flexibility of this kind will be necessary in developing the kind of tools that will enable researchers to use simulation output analysis and results visualisation methods for their models: different output analysis and visualisation techniques may have different requirements for metadata that cannot be comprehensively anticipated at the time simulation output metadata standards are developed.

One of the advantages of social simulation, however, is that there are significant user communities around commonly-used tools such as NetLogo [10] and RePast [11], and that much of the workflow is done using these tools. This creates opportunities to capture metadata manually not only at the time of uploading simulation outputs to a repository, but also during use of the model with appropriate software support, reducing the amount of metadata entered at any one time. Indeed, since activities such as running a simulation model are conducted on a computer, the tools could record associated provenance metadata automatically.

An important question in recording metadata about social simulation outputs is the degree to which the model itself needs to form part of that description. If so, metamodels,

such as MAIA [12], may provide useful vocabularies that can be drawn on in developing simulation output metadata. Bearing in mind the foregoing discussion about the potential for demanding too much from users in the way of manual metadata entry, the benefits of including metadata about the model itself needs to be weighed against the cost in terms of metadata about the outputs that may as a consequence not get recorded.

There is normative pressure for researchers to record metadata about the use of simulation models [e.g. 13, 14], even if only in text form. Much of the information proposed standards or protocols such as these request are amenable to software-facilitated if not automatic capture.

One of the challenges for MIRACLE is to develop a metadata standard for recording simulation output, that allows users (both internal and external to the original project) to explore outputs from agent-based models and build useful queries. The use of tags and keywords could be instrumental in complementing the more formal ontology-based approach to describing these outputs, a prototype of which we have outlined above.

The next steps for MIRACLE involve obtaining requirements for output metadata recording to facilitate development of a standard, and to deploy commonly-used output analysis and visualisation techniques in a web-based platform.

References

[1] P. Fitch, N. Car and T. Smith, "Improving integrated modelling reproducibility and traceability with workflow systems," *Presentation at the International Environmental Modelling and Software Society Congress, 16-20 June 2014, San Diego, CA, USA*.

[2] J. G. Polhill, E. Pignotti, N. M. Gotts, P. Edwards and A. Preece, "A semantic grid service for experimentation with an agent-based model of land-use change," *Journal of Artificial Societies and Social Simulation*, vol. 10, iss. 2, p. 2, 2007.

[3] I. Horrocks, P. F. Patel-Schneider and F. van Harmelen, "From SHIQ and RDF to OWL: The making of a Web Ontology Language," *Journal of Web Semantics*, vol. 1, p. 7-26, 2003.

[4] J. G. Polhill, N. M. Gotts and A. N. R. Law, "Imitative versus nonimitative strategies in a land-use simulation," *Cybernetics and Systems*, vol. 32, iss. 1-2, pp. 285-307, 2001.

[5] J. W. White, A. Rassweiler, J. F. Samhouri, A. C. Stier and C. White, "Ecologists should not use statistical significance tests to interpret simulation model results," *Oikos*, vol. 123, iss. 4, pp. 385-388, 2014.

[6] E. Pignotti, G. Polhill and P. Edwards, "Using provenance to analyse agent-based simulations," In G. Guerrini (ed.) *EDBT '13 Proceedings of the Joint EDBT/ICDT 2013 Workshops*, pp. 319-322.

[7] K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik and J. Zhao, "PROV-O: The PROV Ontology," in T. Lebo, S. Sahoo and D. McGuinness (eds.) *W3C Recommendation 30 April 2013*, http://www.w3.org/TR/prov-o/ (accessed 7 April 2014)

[8] E. Pignotti, P. Edwards, N. Gotts and G. Polhill, "Enhancing workflow with a semantic description of scientific intent," *Journal of Web Semantics*, vol. 9, pp. 222-244, 2011.

[9] P. Edwards, E. Pignotti, C. Mellish, A. Eckhardt, K.Ponnamperuma, T. Bouttaz *et al.*, "Lessons learnt from the deployment of a semantic virtual research environment," *Journal of Web Semantics*, submitted for publication.

[10] U. Wilenski, *NetLogo*. Center for Connected Learning and Computer-Based Modeling, 1999. http://ccl.northwestern.edu/netlogo/

[11] M. J. North, N. T. Collier and J. R. Vos, "Experiences creating three implementations of the Repast agent modeling toolkit," *ACM Transactions on Modeling and Computer Simulation*, vol. 16, iss. 1, pp. 1-25, 2006.

[12] A. Ghorbani, P. Bots, V. Dignum and G. Dijkema, "MAIA: a framework for developing agent-based social simulations," *Journal of Artificial Societies and Social Simulation*, vol 16, iss. 2, p. 9. http://jasss.soc.surrey.ac.uk/16/2/9.html

[13] M. Richiardi, R. Leombruni, N. Saam and M. Sonnessa, "A common protocol for agent-based social simulation," *Journal of Artificial Societies and Social Simulation*, vol. 9, iss. 1, p. 15, 2006.

[14] A. Schmolke, P. Thorbek, D. L. DeAngelis and V. Grimm, "Ecological models supporting environmental decision making: a strategy for the future," *Trends in Ecology and Evolution*, vol. 25, iss. 8, pp. 479-486, 2010.