Transparency of Environmental Computer Models

Martine G. de Vos and **Jan Top** and **Willem Robert van Hage** and **Guus Schreiber** [1]

Abstract. Environmental computer models are considered essential tools in supporting environmental decision making, but their main value is that they allow a better understanding of our complex environment. Despite numerous attempts to promote good modelling practice, transparency of current environmental computer models is limited, which hinders progress in both science and policy making. An important cause is that the structure, meaning and context of environmental computer models is often not clear for other people than the model developers. In the proposed research project we would like to find out whether it is possible to increase the transparency of environmental computer models by making their underlying conceptual model explicit. In preliminary research we identified the following challenges: 1) many model developers are mainly focused on the computational instead of the descriptive aspects of computer models 2) many environmental modellers may not consider the lack of transparency a big problem nor do they see computer scientists as natural partners in cooperation. However, we think that both environmental and computer science could benefit from an interdisciplinary or even a totally integrated approach. We expect that experimenting with tools and methods from computer science could teach us important lessons on the practice of environmental modelling and hopefully guide us to this novel, integrated way of performing e-science.

1 Introduction

1.1 Environmental Computer Models

Current environmental issues have features that distinguish them from traditional scientific problems. They are universal in their scale and long-term in their impact, their mechanisms are complex, variable and not well understood and empirical data are scarce or inadequate [5, 18, 21]. In addition there is an urgent need to find strategies to cope with these issues and political pressure on the research community is high [21].

Environmental computer models are simplified and controllable representations of natural systems, developed by scientists. These models include knowledge and data on the key mechanisms and factors that explain the behaviour of natural systems in a certain context. Although it is hardly possible to validate the results of environmental computer models [13], they are essential tools in supporting environmental decision making by exploring the consequences of alternative policies or management scenarios [5, 18]. They are used to support important political decisions and national investments like the construction of dikes and the design of the future energy system. But the main value of environmental models is that they allow a better understanding of our complex environment [13].

[1] Computer Science, Network Institute, VU University Amsterdam, the Netherlands, email:{Martine.de.Vos—W.R.van.Hage—J.L.Top—Guus.Schreiber}@vu.nl

1.2 Problem Description

Both the developing process and the computer model itself need to be transparent, in order to enable stakeholders and colleague scientists to understand and use environmental computer models. They need to be able to trace model results and insights through the model structure to the underlying choices and assumptions made by the developer [13, 21]. Despite numerous attempts to promote good modelling practice, transparency of current environmental computer models is limited [18, 1] As a consequence: 1) model results and insights may be used in applications without respecting and discussing their underlying choices and assumptions, and 2) learning from model results and insights is difficult, which hinders progress in both science and policy making.

An important cause is that the structure, meaning and context of environmental computer models is often not clear for other people than the model developers [22]. In the development process modellers inevitably make choices on which processes and concepts to include and which to simplify or neglect [5, 21], but they do not make these assumptions explicit. This, in turn, is caused by the lack of short-term incentives for modellers to provide structure, meaning and context to their models, [2, 18, 7] and the size and complexity of their models [9].

2 Related work

Many authors in the field of environmental modelling advocate standardization of the modelling process and information, summarized to as Good Modelling Practice, to enhance transparency of environmental models [14, 5, 17]. The question is whether providing guidelines is sufficient, as the difficulty is not that the elements of Good Modelling Practice are not known or shared, but that the modelling community lacks the urge to act accordingly [1, 18, 16].

In recent years significant progress has been made in the semantic annotation of scientific models, data work flows and publications. In scientific model development ontologies are used to facilitate conceptualization and to achieve shared understanding among model developers and stakeholders [6]. Ontologies are also widely used to semantically annotate scientific models, datasets and publications, i.e., to connect measurements and terms to the identity of observable entities they quantify [11, 19, 8]. A higher level of abstraction that is being investigated is the semantic annotation of scientific practice as a whole. Annotation of work flows supports scientists to integrate and analyse data in a correct and meaningful way [20]. The open provenance model, PROV, developed by the W3C provenance working group [2] helps scientists to document and process provenance information to ensure reproducibility of their analyses [10].

[2] W3C Provenance Working Group,`http://www.w3.org/2011/prov/`

However, in the described annotation methods the models themselves remain largely black-boxes. As a consequence, we may miss out on valuable information on the developers' understanding and interpretation of the system of interest, which is captured in, for example, the used modelling paradigm, the chosen concepts and their interrelations, and the mathematical equations [22].

3 Approach

This research aims to enable developers and stakeholders to cooperate in the development and use of computer models, and to discuss real-world issues not only on the level of model results but also on the level of functioning of the corresponding natural system.

The main central concept of his study is 'transparency'. We define transparency of a computer model as the connections between model concepts, underlying datasets, related publications and knowledge of the model developer. A transparent computer model, with these connections in place, enables peers and stakeholders to 1) understand the knowledge captured in the computer model and 2) to trace back model results and insights through the model structure to this knowledge.

The second important concept of this study is 'conceptual model'. We define the conceptual model of environmental computer models as a knowledge level model [12] containing the concepts that are included, their definitions and their interrelations. The conceptual model represents the basic premises and knowledge about the working of the system being modelled [5] [14].

The main research question of this research is: *Is it possible to increase the transparency of environmental computer models by making their underlying conceptual model explicit?*

3.1 Preliminary results

We did preliminary research on representing the knowledge underlying environmental computer models. In two case studies on existing environmental computer models we manually reconstructed the underlying conceptual model and formally described it in an ontology.

In the first case study [3] we analysed a computational model that determines the energy use by Indian households. The model specifically addresses the socio economic factors influencing energy uses and includes knowledge on consumer behaviour, public health and sustainable development. We used the model documentation, the model source code and personal communication with the model developer to list and define the concepts and their interrelations and represented them in an OWL ontology [3]. We used this ontology in a peer reviewed model evaluation. Scientists representing different disciplines, viz., economics, sustainable development, energy and public health, were asked to determine if the model consisted of the right elements to achieve its goal, or that elements should be added or deleted. They were provided with a visual representation of the ontology (figure 1), i.e. a UML like diagram, and a glossary of the terms in the ontology, as well as the model documentation and source code. We found that the ontology helped these peers to obtain more information on the model and to gain more insight in its structure. However, they lacked time to get a clear overview of the model and were confused by the different sources of information. We concluded that a better balance between different types of model documentation and explicit links between them are needed to really improve the understanding of the model by the peers. An ontology could be useful in bridging the gap between formal documentation, like source code, and documentation in natural language, like reports and papers.

In the second case study [4] we analysed a spreadsheet model that enables policy analyses concerning the Dutch energy system. We studied the design of the tables and the formulas in the spreadsheets [4] and semantically characterized the underlying concepts and their interrelations (figure 2). We represented these as an instantiation of an existing ontology, the OM Ontology for units of Measure and related concepts [15], and verified our findings with the model developers. We found that the both the spreadsheet design and the formulas contain implicit knowledge about the semantics. The main concepts and their interrelations as we identified them in our ontology did not conflict with the developer's views. But we found that representing the conceptual model in an ontology represented a different perspective, as the developers were primarily focussed on the calculation work flow. The developers may see environmental computer models mainly as instruments to perform simulation studies, and therefore focus on the computational aspects, while we see them as tools to communicate scientific knowledge and therefore focus mainly on the descriptive aspects.

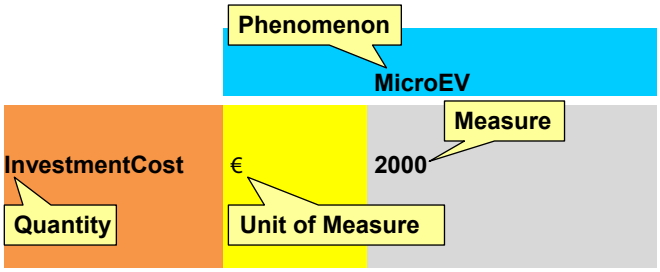

Figure 2. Example, in outline, of the semantic characterization of terms in a spreadsheet table.

3.2 Future work

In future work we intend to perform several case studies on existing environmental computer models. We intend to perform experiments with stakeholders to investigate to what extent reconstructing conceptual models is helpful in understanding and reusing these models. In more detail, we would like to test which form of (visual) presentation of the conceptual model works best to achieve transparency, which aspects of transparency are influenced and to what extent.

The reconstructions in our first case studies were performed manually. During the project we intend to investigate to what extent it is efficient and effective to automatize the process of reconstruction and visual presentation.

We also plan to analyse written (scientific) publications on environmental computer models and the results of their analyses. These publications are often the only way of access to computer models for stakeholders. We would like to find out to what extent it is possible to derive the underlying conceptual model from the written publication and to relate it to the actual content of the computer model.

[3] W3c Web Ontology Language, `http://www.w3.org/TR/owl-features/`

[4] Spreadsheet Examples,`http://semanticweb.cs.vu.nl/edesign/`

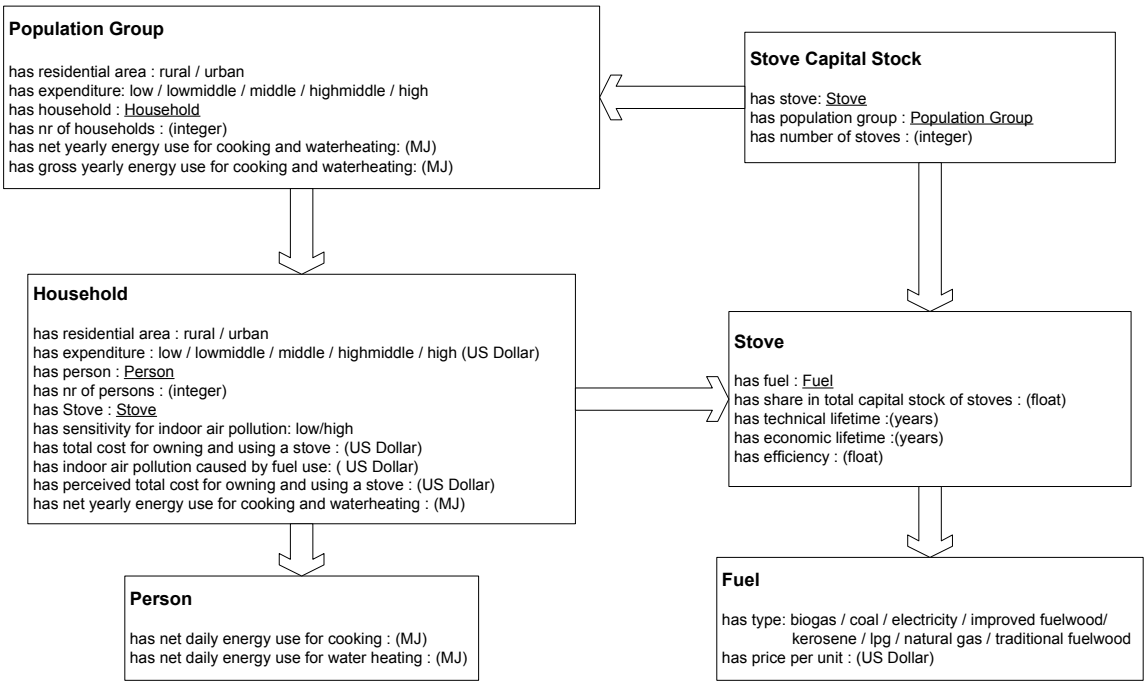

Figure 1. Simplified, visual representation of the ontology of a computational model that determines the energy use by Indian households.

4 Discussion

Our preliminary research gave rise to some additional questions and challenges. Considering our finding that model developers are mainly focused on the computational instead of the descriptive aspects of computer models, we could question whether ontologies are the right method of representing the underlying conceptual knowledge. Analysis and representation of the calculation work flow might also be an option to make the content of a computer model more explicit, but it is not clear to what extent it will contribute to the transparency of computer models. A combined approach is also possible, for example by relating formulas in calculation work flow to concepts in the ontology. An important benefit of ontologies is that they are formal representations and subsequently are amenable to computer processing. (Semi)Automatic analysis of model content and corresponding data and publications could be helpful to achieve transparency of environmental computer models in a more efficient and effective way.

Besides, the request for more transparency is mainly coming from society. Many environmental modellers may not consider the lack of transparency a big problem nor do they see computer scientists as natural partners in cooperation. We think that both environmental and computer science could benefit from an interdisciplinary or even a totally integrated approach. We expect that experimenting with tools and methods from computer science could teach us important lessons on the practice of environmental modelling and hopefully guide us to this novel, integrated way of performing e-science.

5 Matches with other submissions

We see some parallels between our study and submissions 1 and 7. The researchers of submission 1 aim to increase the transparency, which they call scrutability, of autonomous systems. They intend to develop a clear and understandable way to represent formal reasoning models in these systems to humans by translating them into natural language expressions. It would be interesting to compare their and our ways of reconstructing and presenting the conceptual knowledge underlying computer models. Furthermore, their approach could be applicable to the analysis and representation of the calculation work flow in environmental computer models.

The problems concerning computational models in the financial system described by researcher of submission 7 are quite similar to the problems we encounter with environmental computer models. The reliability of these models is questioned and there is a lack of suitable validation/verification techniques. We wonder whether (the lack of) transparency is an issue for these models. Should these models be understandable for non-experts? What type of assumptions and choices are made in these models and to what extent do they influence model results?

REFERENCES

[1] G.a. Alexandrov, D. Ames, G. Bellocchi, M. Bruen, N. Crout, M. Erechtchoukova, A. Hildebrandt, F. Hoffman, C. Jackisch, P. Khaiter, G. Mannina, T. Matsunaga, S.T. Purucker, M. Rivington, and L. Samaniego, 'Technical assessment and evaluation of environmental models and software: Letter to the Editor', *Environmental Modelling & Software*, **26**(3), 328–336, (March 2011).

[2] Nick Barnes, 'Publish your computer code: it is good enough.', *Nature*, **467**(7317), 753, (October 2010).

[3] M.G. De Vos, N Koenderink, B Van Ruijven, and J Top, 'The use of ontologies in peer reviews of Integrated Assessment Models', in *Proceedings of the iEMSs Fifth Biennial Meeting International Congress on Environmental Modelling and Software*, eds., David A. Swayne, Wanhong Yang, Alexey A. Voinov, Andrea Rizzoli, and Tatiana Filatova, pp. 1207–1214, (2010).

[4] M.G. De Vos, Willem Robert Van Hage, Jan Ros, and Guus Schreiber, 'Reconstructing Semantics of Scientific Models : a Case Study', in *Proceedings of the OEDW workshop on Ontology engineering in a data driven world, EKAW 2012*, Galway, Ireland, (2012).

[5] a Jakeman, R Letcher, and J Norton, 'Ten iterative steps in development and evaluation of environmental models', *Environmental Modelling & Software*, **21**(5), 602–614, (May 2006).

[6] S. Janssen, F. Ewert, Hongtao Li, I.N. Athanasiadis, J.J.F. Wien, O. Thérond, M.J.R. Knapen, I. Bezlepkina, J. Alkan-Olsson, a.E. Rizzoli, H. Belhouchette, M. Svensson, and M.K. van Ittersum, 'Defining assessment projects and scenarios for policy support: Use of ontology in Integrated Assessment and Modelling', *Environmental Modelling & Software*, **24**(12), 1491–1500, (December 2009).

[7] Kurt Kleiner, 'Data on demand', *Nature Climate Change*, **1**(April), (2011).

[8] Carla Geovana N. Macário, Sidney Roberto de Sousa, and Claudia Bauzer Medeiros, 'Annotating geospatial data based on its semantics', in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*, p. 81, New York, New York, USA, (2009). ACM Press.

[9] Zeeya Merali, 'Why scientific programming doesn't compute', *Nature*, **467**, 6–8, (2010).

[10] Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, and Jan Van den Bussche, 'The Open Provenance Model core specification (v1.1)', *Future Generation Computer Systems*, **27**(6), 743–756, (June 2011).

[11] Roberto Navigli, Paola Velardi, Alessandro Cucchiarelli, and Francesca Neri, 'Quantitative and Qualitative Evaluation of the OntoLearn Ontology Learning System', in *Proceedings of the 20th international conference on Computational Linguistics*, (2004).

[12] Allen Newell, 'The knowledge level', *Artificial Intelligence*, **18**(1), 87–127, (January 1982).

[13] Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz, 'Verification , Validation , and Confirmation of Numerical Models in the Earth Sciences', *Science*, **263**(5147), 641–646, (1994).

[14] J Refsgaard, 'Modelling guidelinesterminology and guiding principles', *Advances in Water Resources*, **27**(1), 71–82, (January 2004).

[15] H. Rijgersberg, M. Wigham, and J.L. Top, 'How semantics can improve engineering processes: A case of units of measure and quantities', *Advanced Engineering Informatics*, **25**(2), 276–287, (April 2011).

[16] Muir Russell, Geoffrey Boulton, Peter Clarke, David Eyton, and James Norton. The Independent Climate Change E-mails Review, 2010.

[17] Edward J. Jr. Rykiel, 'Testing ecological models: the meaning of validation', *Ecological Modelling*, **90**, (1996).

[18] Amelie Schmolke, Pernille Thorbek, Donald L DeAngelis, and Volker Grimm, 'Ecological models supporting environmental decision making: a strategy for the future.', *Trends in ecology & evolution*, **25**(8), 479–86, (August 2010).

[19] Tony C Smith and John G Cleary, 'Automatically linking MEDLINE abstracts to the Gene Ontology', in *Proc. ISMB 2003 BioLINK Text Data Mining SIG*, pp. 1–4, (2003).

[20] Jacek Sroka, Jan Hidders, Paolo Missier, and Carole Goble, 'A formal semantics for the Taverna 2 workflow model', *Journal of Computer and System Sciences*, **76**(6), 490–508, (September 2010).

[21] Jeroen P. van der Sluijs, 'A way out of the credibility crisis of models used in integrated environmental assessment', *Futures*, **34**(2), 133–146, (March 2002).

[22] Ferdinando Villa, Ioannis N. Athanasiadis, and Andrea Emilio Rizzoli, 'Modelling with knowledge: A review of emerging semantic approaches to environmental modelling', *Environmental Modelling & Software*, **24**(5), 577–587, (May 2009).