Interpreting environmental computational spreadsheets

Martine de Vos

Computer Science, Network Institute, VU University Amsterdam, the Netherlands
Martine.de.Vos@vu.nl

Abstract. Environmental computational spreadsheets are important tools in supporting decision making. However, as the underlying concepts and relations are not made explicit, the transparency and re-use of these spreadsheets is severely limited. The goal of this project is to provide a semi-automatic methodology for constructing the underlying knowledge level model of environmental computational spreadsheets. We develop and test this methodology in a limited number of case studies. Our methodology combines heuristics on spreadsheet layout and formulas, with existing methods from computer science. We evaluate our constructed model with both the original developers and their peers.

1 Problem Statement

Current environmental issues, like climate change and biodiversity loss, are universal in their scale and long-term in their impact, their mechanisms are complex, and empirical data are scarce [1–3]. In addition there is an urgent need to find strategies to cope with these issues, and political pressure on the research community is high [3]. Environmental computer models are considered essential tools in supporting environmental decision making by exploring the consequences of alternative policies or management scenarios [1, 2].

Environmental computer models are mainly developed and used by domain scientists and typically implemented as spreadsheets, Fortran programs or in MatLab. These domain scientists have a knowledge level model [4] in their minds containing the important concepts in their domain, and corresponding definitions and interrelations. In the model development process (figure 1) they inevitably make choices about which entities and processes they should include to describe their study area, and how these should be translated and implemented in their computer model. In this way their knowledge model is implicitly included in the computer model, as it is reflected in, for example, the used modelling paradigm, the model structure, the chosen concepts and their interrelations, and the mathematical equations [5].

It is hardly possible to obtain the knowledge level model from the domain scientists themselves. They may give a limited textual explanation about their ideas and choices in their publications, but they rather focus on the computational side of modeling [6]. In fact, they may not even be aware of the knowledge

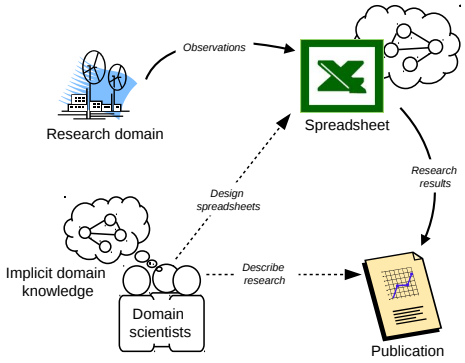

Fig. 1. Rough schematic overview of the current practice of development and use of scientific spreadsheets by domain scientists

level model in their mind [7]. The knowledge level model is, however, essential to understand the meaning and context of the results and insights generated with the computer model. As a consequence, it is hard to make efficient and effective use of environmental computer models by other people than the original developers [6].

The focus of this research is on environmental computer models that are implemented as spreadsheets, from now on called 'environmental computational spreadsheets'. Spreadsheets are widely used by domain scientists to store and manipulate quantitative data from their research projects [8, 9]. A drawback of current spreadsheets is that their free format leads to both complex layout of tables, and sloppy or limited specification of the semantics of the data and calculations [10, 11]. The goal of this project is therefore to provide a methodology for making the underlying knowledge level model of environmental computational spreadsheets explicit. Ideally the various elements in the research process, i.e. observational data, spreadsheet and publications, could be connected to each other through this explicit knowledge level model.

2 Relevancy

Results of this research could enable peers to discuss and assess the scientific quality of environmental computational spreadsheets and to reuse corresponding results and insights. This could contribute to both scientific cooperation and progress, and reliable environmental decision making.

Our research is focused on spreadsheets from the domain of environmental science. However, scientists from other domains may have a similar way of designing and using their spreadsheet models as environmental scientists. We therefore think that the methods and insights from this study might also be applied to spreadsheets from other domains, provided that these spreadsheets contain both domain knowledge and quantitative data.

3 Related Work

Many authors in the field of environmental science advocate standardization of the modelling process, summarized to as 'Good Modelling Practice', to enhance transparency of environmental computer models [12, 1, 13]. Similarly, several studies in computer science, especially in the field of software engineering, suggest how scientific software development could benefit from, for example, clear documentation, relevant training options for scientists and publication of source code [14, 15, 17]. The suggested procedures and guidelines will likely yield more reliable software. However, to guarantee more reliable science, the knowledge included in that software should also be taken into account.

In recent years significant progress has been made in the semantic annotation of scientific models, data sets, and publications. Many tools and techniques are avaliable to connect measurements and terms to the identity of observable entities they quantify [18–21]. A higher level of abstraction that is being investigated is the semantic annotation of scientific practice as a whole. The open provenance model, PROV, [1] helps scientists to document and process provenance information to ensure reproducibility of their analyses [22]. Furthermore, in several scientific disciplines workflow systems [23, 24] are used to integrate and analyse data in a correct and meaningful way.

Several tools and techniques can be used to annotate tabular data. The Data Cube vocabulary [2], for example, provides a means for publishing statistical data as linked data with associated metadata in order to support interpretation and reproducibility. Existing conversion systems like RDF123 [11] and XLWrap [25] allow mapping information from spreadsheets to RDF. And some tools, like Rightfield [8] and Anzo [3], allow the direct annotation of data inside spreadsheet tables.

4 Research Question(s)

In the above described annotation methods the spreadsheets 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. However, related work also shows that there are plenty solutions to the issue of representing scientific tabular data. As such these studies provide useful tools and information that can be used as a starting point for present study.
The general research question we wish to answer in our study is the following:
To what extent can the underlying knowledge level model of an environmental computational spreadsheet be made explicit?
We refine this question into two more specific subquestions.

1. *How can the underlying knowledge level model of an environmental computational spreadsheet be adequately described?*

[1] W3C Provenance Working Group, http://www.w3.org/2011/prov/
[2] Data Cube, http://www.w3.org/TR/vocab-data-cube/
[3] Anzo, http://www.cambridgesemantics.com

An adequate description of the underlying knowledge level model is defined as a description that
  - agrees with the views of the original developers of the spreadsheets.
  - can be understood and applied by the original developers of the spreadsheets and their peers.
  - allows representation of domain concepts, their hierarchical and property relations, and the computational relations that exist between these concepts

2. *What are the requirements for a methodology for constructing the underlying knowledge level model of an environmental computational spreadsheet?*

5 Hypotheses

When we apply our methodology to an environmental computational spreadsheet, we expect that the resulting constructed knowledge level model is an adequate description of the underlying knowledge level model.

6 Preliminary results

We did two case studies on an existing environmental computer model,i.e., a spreadsheet model that enables policy analyses concerning the Dutch energy system .

In the first case study [7] we manually analyzed the design of the tables and the formulas in the spreadsheets [4]. We semantically characterized the underlying concepts and their interrelations (figure 2) and represented these as an instantiation of an existing ontology, the OM Ontology for units of Measure and related concepts [10]. The main concepts and their interrelations as we identified them in our resulting ontology did not conflict with the developer's views. However, we also discovered that the developers see their models mainly as instruments to perform simulation studies, and therefore focus on the computational aspects.

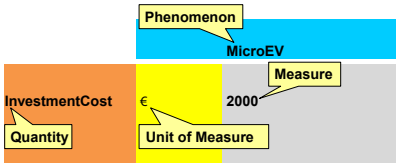

Fig. 2. Example, in outline, of the semantic characterization of terms in a spreadsheet table.

In the second case study [26] we combined automatic and manual methods to analyze the calculation procedures in the spreadsheets. This resulted in a huge

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

network of interconnected spreadsheet cells (figure 3). We used network analysis to determine which nodes in the graph are the most important, and manually connected these in a simplifed calculation workflow.

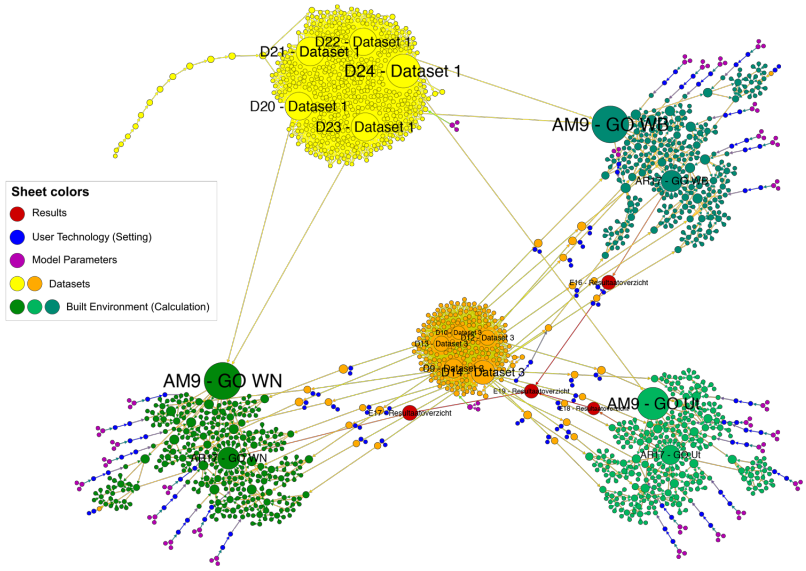

Fig. 3. Network of spreadsheet cells connected through formulas.

7 Approach

In this project we aim at developing a methodology for semi-automatic construction of the underlying knowledge model of an environmental computational spreadsheet. As described above, there are no similar studies on this topic, nor is it possible to access the knowledge level model in the minds of the original developers of environmental computational spreadsheets. We therefore consider it not feasible to set up a study based on quantitative experiments. Instead we choose an approach based on the analyses of a limited number of case studies, and as a consequence, our research has an exploratory character.

Our case studies are all scientific spreadsheet models of existing research projects from the domain of environmental science. We have access to the actual spreadsheets and corresponding datasets, as well as to the publications describing the models and analyses. Furthermore, we have personal contact with the model developers and users.

We develop our methodology based on the in-depth, qualitative analysis of one case study. We will manually analyze the layout of the spreadsheet tables, as well as the formulas connecting the spreadsheet cells. We determine to what extent the observed patterns provide insight in the semantics of the content of the tables, and record our findings in heuristics. Spreadsheet terms can be matched

automatically with concepts of external vocabularies on domain concepts, and on quantitative tabular data. We combine this matching with our layout heuristics to recognize the concepts in the spreadsheets and their interrelations. In addition, we will automatically trace the dependencies between spreadsheet cells through formulas and analyze the resulting networks using techniques for network analysis. We combine these analyses with our heuristics on formulas to construct the calculation workflow in the spreadsheets.

Research question 1 is studied by focusing on the performance of our method in each case study. The different steps in our methodology of constructing the knowledge level model are performed manually by the original developers, and their results are compared with results from our semi-automatic method. We test the applicability of the constructed model by using it to connect concepts from the spreadsheets, with concepts from corresponding publications, or data sets. In a separate user study we will test to what extent peers are able to understand and apply the constructed knowledge level model.

Research question 2 is studied by focusing on the different techniques that are used to describe the knowledge level model. The use of external vocabularies is evaluated by determining how many of the spreadsheet terms could be matched, and how relevant these matches are. We also determine which properties of these vocabularies influence this matching. The use of network analysis techniques is evaluated by determining to what extent these techniques are able to recognize the important variables, as indicated by the original developers, in the calculation workflow. We determine which properties of the spreadsheets influence the performance of our method.

8 Evaluation plan

In order to test our hypothesis we will formulate measurable definitions on what it means for original developers and peers to understand and apply the constructed knowledge level model. Possible indicators we could use are, for example,

- the number of concepts, relations and variables that occur both in the constructed model and in the manual analysis of the original developers.
- the number of connections that can be made from the spreadsheet to corresponding publications and datasets.

9 Reflections

We think our approach is likely to succeed as it is targeted at existing environmental computational spreadsheets. We expect that studying the patterns in these spreadsheets will provide us useful insights on environmental modeling. We also see several promising external developments. Firstly, there is a growing awareness of both the importance of open source code and data, and the importance of methods to provide corresponding credits to modelers and data

providers. Besides, there is an increasing availability of external domain vocabularies.
This PhD research is now at the half way stage. Current work is an extension of our first case study (section 6) and involves the development of a semi-automatic method for defining the concepts and interrelations in spreadsheets. We use the external vocabularies AGROVOC [27] and OM[10], to map and categorize the spreadsheet terms. The plan for the near future is to continue the work of our second case study by developing a semi-automatic method for the construction of the calculation workflow.

Acknowledgements

This publication was supported by the Data2Semantics project in the Dutch national program COMMIT. Guus Schreiber and Jan Top (supervisors), Paul Groth and Lora Aroyo are acknowledged for providing useful comments and suggestions.

References

1. Jakeman, a., Letcher, R., Norton, J.: Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software **21**(5) (May 2006) 602–614
2. Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V.: Ecological models supporting environmental decision making: a strategy for the future. Trends in ecology & evolution **25**(8) (August 2010) 479–86
3. van der Sluijs, J.P.: A way out of the credibility crisis of models used in integrated environmental assessment. Futures **34**(2) (March 2002) 133–146
4. Newell, A.: The knowledge level. Artificial Intelligence **18**(1) (January 1982) 87–127
5. Villa, F., Athanasiadis, I., Rizzoli, A.E.: Modelling with knowledge: A review of emerging semantic approaches to environmental modelling. Environmental Modelling & Software **24**(5) (May 2009) 577–587
6. De Vos, M., Janssen, S., Van Bussel, L., Kromdijk, J., Van Vliet, J.V., Top, J.L.: Are environmental models transparent and reproducible enough ? In Wongsosaputro, J., Pauwels, L., Chan, F., eds.: Proceedings of 19th International Congress on Modelling and Simulation. (2011) 2954–2961
7. De Vos, M., Van Hage, W.R., Ros, J., Schreiber, A.: 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)
8. Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J.L., du Preez, F., Goble, C.: RightField: embedding ontology annotation in spreadsheets. Bioinformatics (Oxford, England) **27**(14) (July 2011) 2021–2
9. Rocha Bernardo, I., Mota, M.S., Santanchè, A.: Extracting and Semantically Integrating Implicit Schemas from Multiple Spreadsheets of Biology based on the Recognition of their Nature. Journal of Information and Database Management **4**(2) (2013) 104–113

10. Rijgersberg, H., Wigham, M., Top, J.L.: How semantics can improve engineering processes: A case of units of measure and quantities. Advanced Engineering Informatics **25**(2) (April 2011) 276–287
11. Han, L., Finin, T., Parr, C., Sachs, J., Joshi, A.: RDF123 : From Spreadsheets to RDF. In: The Semantic Web-ISWC 2008, Springer Berlin Heidelberg (2008) 451–466
12. Refsgaard, J.C.: Modelling guidelinesterminology and guiding principles. Advances in Water Resources **27**(1) (January 2004) 71–82
13. Rykiel, E.J.J.: Testing ecological models: the meaning of validation. Ecological Modelling **90** (1996)
14. Hannay, J.E., Macleod, C., Singer, J., Langtangen, H.P., Wilson, G.: How Do Scientists Develop and Use Scientific Software ? In: Proceedings of the 2009 ICSE workshop on Software Engineering for Computational Science and Engineering, IEEE Computer Society (2009)
15. Segal, J., Morris, C.: Developing scientific software, Part 2. IEEE software **26**(1) (2009) 79
16. Merali, Z.: Why scientific programming doesn't compute. Nature **467** (2010) 6–8
17. Bizer, C., Berlin, F.U., Seaborne, A., Labs, H.p.: D2RQ Treating Non-RDF Databases as Virtual RDF Graphs. In: Proceedings of the 3rd International Semantic Web Conference (ISWC2004). (2004)
18. Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F.: Quantitative and Qualitative Evaluation of the OntoLearn Ontology Learning System. In: Proceedings of the 20th international conference on Computational Linguistics. (2004)
19. Ngomo, A.c.N., Auer, S.: LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, Volume Three., AAAI Press, (2011) 2312–2317
20. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk A Link Discovery Framework for the Web of Data. (2009)
21. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., den Bussche, J.V.: The Open Provenance Model core specification (v1.1). Future Generation Computer Systems **27**(6) (June 2011) 743–756
22. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M.B., Lee, E.a., Tao, J., Zhao, Y.: Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience **18**(10) (August 2006) 1039–1065
23. Sroka, J., Hidders, J., Missier, P., Goble, C.: A formal semantics for the Taverna 2 workflow model. Journal of Computer and System Sciences **76**(6) (September 2010) 490–508
24. Langegger, A., Wolfram, W.: XLWrap Querying and Integrating Arbitrary Spreadsheets with SPARQL. (2009) 359–374
25. De Vos, M., van Hage, W.R., Wielemaker, J., Schreiber, A.: Knowledge Representation in Scientific Models and their Publications : a Case Study. In: Proceedings of K-CAP 2013 Knowledge Capture Conference, Banff, Canada (2013) 1–2
26. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering Thesauri for New Applications : the AGROVOC Example. Journal of Digital Information **4**(4) (2004) 1–15