Knowledge Representation in Scientific Models and their Publications: a Case Study

Martine G. de Vos
Computer Science, Network Institute
VU University Amsterdam
de Boelelaan 1081a, 1108 HV
Amsterdam, The Netherlands
Martine.de.Vos@vu.nl

Willem Robert van Hage
Computer Science, Network Institute
VU University Amsterdam
de Boelelaan 1081a, 1108 HV
Amsterdam, The Netherlands
W.R.van.Hage@vu.nl

Jan Wielemaker
Computer Science, Network Institute
VU University Amsterdam
de Boelelaan 1081a, 1108 HV
Amsterdam, The Netherlands
J.Wielemaker@vu.nl

Guus Schreiber
Computer Science, Network Institute
VU University Amsterdam
de Boelelaan 1081a, 1108 HV
Amsterdam, The Netherlands
Guus.Schreiber@vu.nl

ABSTRACT

Environmental computer models are simplified representations of natural systems and are essential tools in studying our complex environment. These models and simulation results are described in publications. Both the model itself and the publication are needed to get a complete picture of the underlying research project. However, the publication is often used as the only source of information. In this study we explore to what extent and by what means the links between a publication and the associated computational model can be made explicit. We perform a semi-automatic analysis of the spreadsheets and report of an existing research project used to inform the Dutch government about energy policy. We find that the report chapter and spreadsheets agree at the conceptual level, but the difference in abstraction level makes it difficult, however, to create explicit links between them. We discuss how an intermediate conceptual model could be used to describe concepts and variables in the underlying research project, and to enable links with both spreadsheet and publication.

Keywords

conceptual modeling, spreadsheets, workflows, ontology, corpus statistics, network analysis

1. INTRODUCTION

Environmental computer models are simplified and controllable representations of natural systems, developed by scientists and typically implemented as spreadsheets, Fortran programs or in MatLab. Such models include detailed knowledge and data on the key mechanisms and factors that explain the behaviour of natural systems in a certain context. Computational models are essential tools in supporting environmental decision making by exploring through simulations the consequences of alternative policies or management scenarios [9, 10].

The models and the performed simulations are typically described in a paper or a report. This publication does not serve as documentation, but rather provides peers and stakeholders with an explanation of the underlying concepts and an interpretation of the simuation results. The model and associated publication have different goals and focus, but are both needed to get a complete picture of the underlying research. Although more and more computational models and associated data become publicly accessible, the models themselves are often too big or too complicated for people to understand easily. In practice, the publication serves as single source of information on the underlying research project. However, it would be desirable if the publication was linked to key elements of the relevant computational model. In this way, the publication can provide insight into the model structure and calculation of results. This insight is needed to enable peers and stakeholders to understand and use computational models and associated data. [1]

The objective of this paper is to explore to what extent and by what means the links between a publication and the associated computational model can be made explicit. We perform a case study on the spreadsheets and report of an existing research project. We use corpus statistics to find out which terms are important in the spreadsheets and to what extent these are included in the report. We use network analysis of the spreadsheet and manual analysis of the report text to compare the included variables and formulas. As result of this explorative study we expect to find which elements in the report and spreadsheets can be linked directly , and what is needed to actually implement

[1]Peers and stakeholders need to know which concepts are included in the model and how these are interrelated. They need to be able to trace model results from the publication through the formulas to the model concepts.

these links.

The outline of this paper is as follows: First we describe related research in section 2. Next we explain the setup of the case studyin section 3, and in section 4 we report on the results. Finally, in section 5 we evaluate the results and discuss options for future research

2. RELATED WORK

In the last decade much research has been done on connecting scientific data and analyses to facilitate the exchange and reuse of digital knowledge. Publishing models and datasets as linked data [2] enables connections with related data sets and concepts. 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. In several scientific disciplines workflow systems (e.g., [7];[5]; [1]) are used to support data-intensive research by connecting the various data manipulation, analysis and reportation steps. Such reconstruction of workflows facilitate the documentation and reproducibility of whole experiments, but they require all the data and methods to be executed a-priori in the system. Post-hoc reconstruction of the provenance of existing documents and data is much more complicated and only beginning to be adressed [6]. In this era of eScience it becomes increasingly important to describe and annotate data and models, not only to allow reproducible research, but also to allow understanding and therefore enhance the exchange of knowledge [4].

3. CASE STUDY

Our case study is a scientific computer model for energy policy analysis [3], developed by the PBL Netherlands Environmental Assessment Agency (PBL) and the Energy Research Centre of the Netherlands (ECN). The computer model consists of a set of spreadsheets. The associated publication is a PBL research report [8] used to inform the Dutch government about consequences of different strategies in energy policy.

We include all the spreadsheets in the our analysis. The text we select for our investigation is a chapter from the report describing the computer model and the experimental results. We determine which terms are mentioned in the spreadsheets and compare these to the terms included in the report chapter. We reconstruct calculation procedures from the storyline in the chapter text and investigate to what extent these agree with the calculation procedures included in the spreadsheets.

3.1 Analysis of concepts

To analyze which terms are included in the report chapter, we compare the terms in the chapter text with terms in a corpus. Our corpus consists of three reports, viz., the complete PBL research report and two research reports of two other govermental agencies in different fields, i.e. the Netherlands Institute for Social Research and the CPB Netherlands Bureau for Economic Policy Analysis. We compute the term frequency-inverse document frequency (tf-idf), to evaluate the importance of a term to the report chapter relative to its importance to the corpus: We split all reports into pages and compute the tf-idf as defined in Eq.(1):

$$\text{tf} - \text{idf} = \log(1 + \text{count}(t, page)) \times \log \frac{|corpus|}{\text{countpage}(t, corpus)}$$

where count*(t,page)* refers to the frequency of term t on a page, $|corpus|$ refers to the number of pages in the corpus and countpage*(t,corpus)* refers to the number of pages in the corpus that contain the term t. By selecting the terms with a tf-idf score greater than 0, we retrieve a set of terms that is deprived from common words which have no importance to the chapter text.

We obtain the terms used in the computer model by automatically retrieving the labels, i.e., the cells of type *string*, from all the spreadsheets in the model, which we then tokenize and count.

Subsequently, we determine the nature and fraction of spreadsheet terms that are included in the chapter text and the nature and fraction of spreadsheet terms that are left out. We also determine the nature of the report chapter terms that do not occur in the spreadsheets. As we do not have a Dutch dictionary available containing terms specific to the domain of energy science, we can not lemmatize the terms from the spreadsheet and chapter before the analysis. The results are manually corrected afterwards, to make sure that not only exact matches of terms are considered but also inflicted forms.

3.2 Analysis of calulation procedures

The chapter of the research report describes results of model analyses and the way these are produced. These descriptions are straightforward, which enables us to, to a certain extent, reconstruct the corresponding formulas from the storyline in the text. We retrieve the formulas from the report sections, analyze how they are related through their in and output and reconstruct chains of formulas. We also analyze how they are related to concepts (domain specific terms) in the text. As an example we explain, step by step, the analyis of the formula of one of the lines in the chapter section:
Report text *"The overall amount of the mentioned items is 45 to 70 Mton"*

- the term *"overall amount of 45 to 70 Mton"* represents the output of a sum
- the term *"overall amount 45 to 70 Mton"* refers to the total of remaining greenhouse emissions mentioned earlier in the same pararagraph
- the term *"mentioned items"* represents the input of the abovementioned sum
- the term *"mentioned items"* refers to a list of greenhouse gases emissions per economic sector mentioned earlier in the same pararagraph

We retrieve calculation procedures from the spreadsheet by automatically tracing the dependencies between individual

[2]Data Cube, `http://www.w3.org/TR/vocab-data-cube/`
[3]Edesign, `http://www.pbl.nl/e-design/`

cells through the included formulas. Analyzing the dependencies between all the cells in the spreadsheet would produce huge networks with thousands of nodes and edges. From the point of view of feasibility, and as proof op principle, we select one of the cells on the sheet that is presenting the overall results of the model. Subsequently, we use network analysis to determine which nodes in the graph are the most important. We use a product of normalized *pagerank* and normalized *betweennes*, which represent respectively the nodes with the most central positions and the nodes which act as gatekeepers. The resulting "most important" nodes we connect in a reconstructed, simplified calculation workflow.

Subsequetly, we determine which variables and relations from the chapter text are found in the spreadsheets, and which are left out. We also determine which variables and relations from the reconstructed spreadsheet workflow are included in the chapter text, and which are not present.

4. RESULTS

4.1 Term analysis

The total number of terms retrieved from the spreadsheet was 852. A fraction of 0.39 of these terms could be found in the chapter text. From the top 30 most frequent terms in the spreadsheets 23 terms were present in the chapter text (table 1).

Many frequent terms in the spreadsheets represent units of measure, e.g., *pj*, *twh* and *mton* (table 1), or refer to model or parameter settings, e.g., *unlimited* (table 3). Except for the units of measure represented in table 1, no units were found in the chapter text, irrespective of their spelling or the use of their full names. Some of the model settings refer to the research design, e.g., *pessimistic* and *optimistic* (table 1) which refer to the used scenario , and several of these terms are also present in the chapter text. The model and parameter settings that refer to calculation processes were not found in the chapter text.

The chapter text consists mainly of domain specific terms (table 2) and terms representing high level quantifications or qualifications, e.g., *energy demand* and *clean* (table 4). Domain specific terms were frequently present in the spreadsheets (table 1, 2), but the high level terms were scarce.

The terms found in both spreadsheets and chapter text are semantically equal, i.e., they refer to the same concepts. Some terms in the chapter or spreadsheet may seem ambiguous, but their use and meaning in the context of the spreadsheets and report chapter is very clear. For example, the term *unlimited* (table 3) is a parameter setting in the spreadsheets, which refers to the contribution of individual technologies to the total energy demand, while the term *limited* (table 4) is a qualitative measure in the chapter text, which refers to the availability of stocks of energy carriers. And the term *clean* (table 4) always refers to energy production with no or little greenhouse gas emissions.

Table 1: Top ten terms found in both spreadsheets and chapter text; top ten according to spreadsheet rank

term	chapter tf-idf	spreadsheet count	spreadsheet rank
pj *(peta joule)*	9.8	3667	1
pessimistic	5.6	1057	3
optimistic	7.5	1056	4
twh *(tera watt hour)*	10.8	828	5
heat	10.3	758	6
mton *(mega ton)*	27.3	717	7
biomass	40.3	667	8
km *(kilometer)*	3.5	495	10
co2	38.5	366	13
natural gas	6.1	365	14

Table 3: Top ten terms only found in spreadsheets

term	spreadsheet count	spreadsheet rank
middle	1071	2
billion	496	9
m *(meter)*	383	12
unlimited	363	15
geothermal energy	353	17
lt *(low temperature heater)*	216	28
ht *(forced-air heater)*	216	27
sht *(wood-burning heater)*	208	33
wet	208	32
steal	199	37

Table 2: Top ten terms found in both spreadsheets and chapter text; top ten according to chapter tf-idf

term	chapter tf-idf	spreadsheet count	spreadsheet rank
biomass	40.3	667	8
co2	38.5	366	13
technologies	33.0	3	457
supplies *(stocks)*	32.6	4	419
elektricity	31.0	210	30
hydrogen	30.1	261	21
supply *(availability)*	29.3	23	175
usage	27.7	35	143
mton *(mega ton)*	27.3	717	7
emissisions	26.9	9	303

We found that less than half of the terms in the spreadsheets matched terms in the publication at a lexical level, but there was much more overlap at the conceptual level. (table 5). The chapter text was not confined by the use of abstract or aggregate concepts, but used less than half of the subclasses that were present in the spreadsheets.

4.2 Calulation analysis

In the chapter text 54 different variables were found. A fraction of 0.57 could be traced to cells in the spreadsheets. Most of these were found on the same three spreadsheets, that summarize model settings and results. The variables

Table 4: Top ten terms only found in chapter text

term	chapter tf-idf
energy demand	34.0
capture	31.4
decrease	29.7
clean	25.1
reference scenario	24.3
alternatives	22.9
feasibility	22.7
elektrification	20.9
building blocks	20.9
limited	19.0

Table 5: Number of terms per concept in spreadsheet and chaptertext

Superclass	Nr. of Subclasses		
	in spreadsheet	*in chapter*	*overlap*
Technology	78	31	27
Sector	37	26	19
Supply *(stock)*	24	13	11
Biomass	17	2	2

that could not be found in the spreadsheets were mainly aggregate variables, for example, *total energy demand* and *costs of technologies*. Though, the different components of these aggregate variables, for example, *energy demand per subsystem* and *cost per technology*, were all present in the spreadsheets. About one third of the relations between the chapter variables could be found as direct relations between cells in the spreadsheets. These consisted mainly of relations between *total emissions in the netherlands* and textitemissions of the different economic secors.

Figure 1 shows the 1796 spreadsheetcells that are, through formulas, connected to cell "E19 - Results", which represents one of the end result in the model *total of greenhouse gas emissions in the built environment*. Only 19 cells were characerized as important "central" and "gatekeeper" variables in this network. These 19 variables could be connected, directly and indirectly, in a simplified reconstruction of the calculation workflow of the end result (figure 2). When we compared the reconstructed workflow with the chapter text, we could find the endresult variable, as well as a number of the *used technologies*. The chapter mentioned also an aggregate variable *heat demand built environment* which is not as such present in the workflow, but its components are. None of the other variables and none of the relations in the reconstructed workflow could be found in the chapter text.

5. DISCUSSION

Results of the terms analysis showed that the chapter and the spreadsheets use the same concepts. But the chapter typically focuses on the super and aggregate classes, while in the spreadsheets the low-level classes of the same concepts are more frequent. The calculation procedure of model results as described in the report chapter gives a correct, but incomplete and very general outline of the workflow included in the spreadsheets. The chapter describes many aggregate or abstract variables which were not found as such in the spreadsheets, while the component variables from the spreadsheets were not present in the chapter.

As the report chapter and spreadsheets agree at the conceptual level, linking the two items seems appropriate. The difference in abstraction level makes it difficult, however, to create explicit links between elements in the report and the spreadsheets. An intermediate conceptual model is needed to connect the high level concepts and variables from the report with the low level concepts and variables from the spreadsheets.

Such a conceptual model would ideally exist of two layers: an ontology describing the used concepts and their hierarchical and property relations, and a workflow model describing the research variables and their connections through formulas (figure 3). The concepts and variables present in the conceptual model could then directly be linked to those in the spreadsheet and publication. There should also be links between the two layers within the conceptual model, as concepts in the ontology are usualy related to variables in the workflow model. For example, properties of concepts are often used as variables in formulas.

This study and previous work on the same case [3] could provide the first steps towards the construction of the described conceptual model. In the previous paper we manually reconstructed an ontology of the underlying research from the spreadsheet implementation. We used an existing ontology on units of measurements to characterize and formally describe the terms in the spreadsheet and we recorded our approach in heuristics. We found that both the design of the spreadsheet and the formulas contain implicit but valuable information on the underlying concepts. The tracing of dependencies between cells, and the subsequent network analysis and manual reconstruction, as presented in this study, could be used to construct a workflow model. Although many of the actions performed in these explorative studies are performed manually, we think that there are several opportunities for automation.

6. ACKNOWLEDGEMENTS

We wish to thank PBL researchers Jan Ros en Jeroen Peters for providing us with their model and data, and our colleagues Michiel Hidebrand and Jan Top for their useful comments. This publication was supported by the Data2Semantics project in the Dutch national program COMMIT.

References

[1] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, and C. Goble. Why linked data is not enough for scientists. *Future Generation Computer Systems*, Aug. 2011.

[2] C. Bizer, T. Heath, and T. Berners-lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 2009.

[3] M. De Vos, W. R. Van Hage, J. Ros, and G. 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.

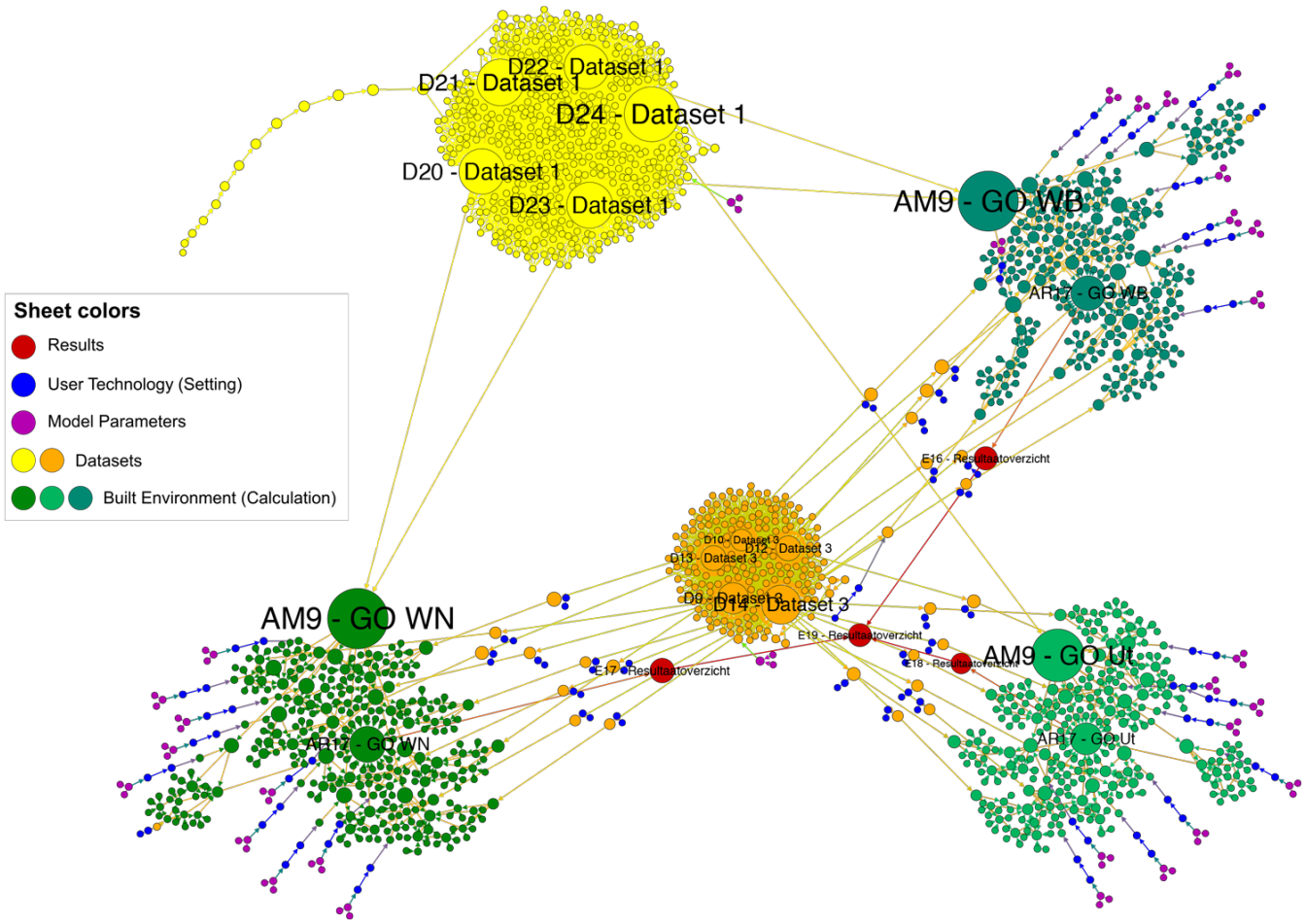

Figure 1: Network analysis of all spreadsheet cells connected to an end result of the computational model. The size of the circles indicates the importance of the corresponding cell as "central" and "gatekeeper" node.

[4] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm; Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[5] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. a. Lee, J. Tao, and Y. Zhao. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065, Aug. 2006.

[6] S. Magliacane. Reconstructing Provenance. In *ISWC 2012*, pages 399–406, Berlin/Heidelberg, 2012. Springer-Verlag.

[7] P. Missier, S. Soiland-reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble. Taverna , reloaded. In *Scientific and Statistical Database Management*, Berlin/Heidelberg, 2010. Springer.

[8] PBL Planbureau voor de Leefomgeving and Energieonderzoek Centrum Nederland. Naar een schone economie in 2050: routes verkend. Technical report, Netherlands Environmental Assesment Agency, Den Haag, 2011.

[9] A. Schmolke, P. Thorbek, D. L. DeAngelis, and V. Grimm. Ecological models supporting environmental decision making: a strategy for the future. *Trends in ecology & evolution*, 25(8):479–86, Aug. 2010.

[10] F. Villa, I. N. Athanasiadis, and A. E. Rizzoli. Modelling with knowledge: A review of emerging semantic approaches to environmental modelling. *Environmental Modelling & Software*, 24(5):577–587, May 2009.

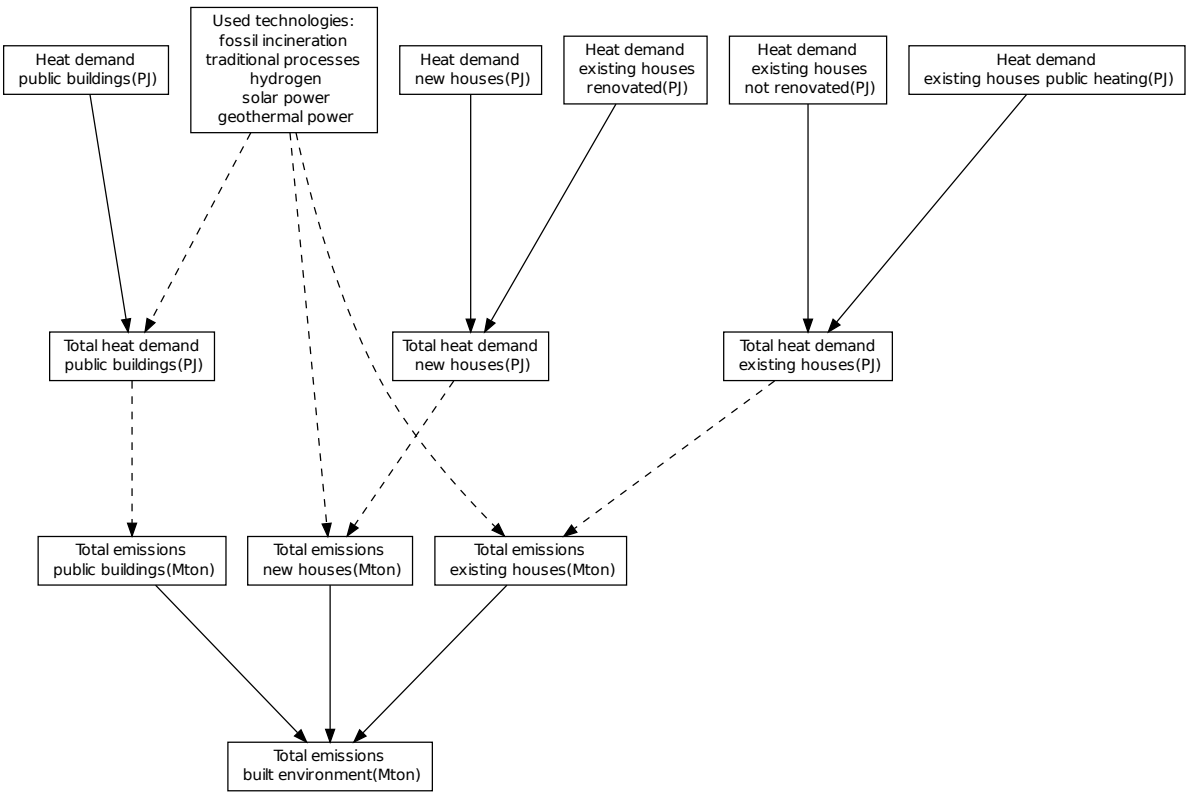

Figure 2: Reconstructed, simplified calculation workflow based on the results of the network analysis in figure 1. Repeating variables are excluded and the different technology choices are represented as one variable.

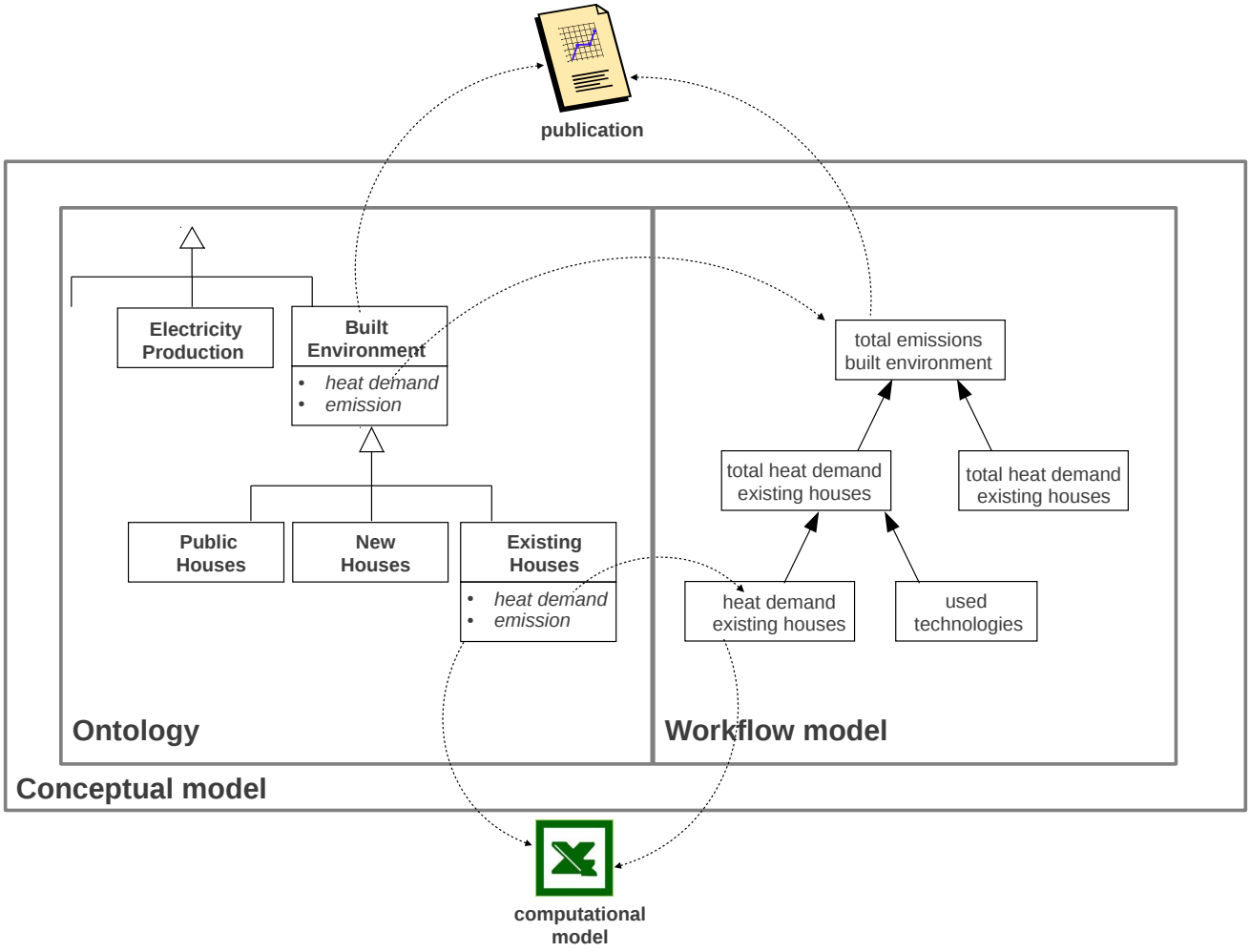

Figure 3: Schematic design of an intermediate conceptual model connecting concepts and variables from the spreadsheet and publication.