Reconstructing Semantics of Scientific Models: a Case Study

Martine de Vos[1], Willem Robert van Hage[1], Jan Ros[2], and Guus Schreiber[1]

[1] Computer Science, Network Institute, VU University Amsterdam, the Netherlands
{Martine.de.Vos|W.R.van.Hage|Guus.Schreiber}@vu.nl
[2] PBL Netherlands Environmental Assessment Agency, Bilthoven, the Netherlands

Abstract. Spreadsheets are frequently used by scientists to store and analyze research data. To enable integration and reusability of scientific spreadsheet data it is important to explicate the underlying concepts and relations. In this paper we explore to which extent the conceptual model of a research project can be recognized in its spreadsheet implementation. We perform a manual analysis of spreadsheets of existing research from the domain of environmental science. We formally describe the semantics of the spreadsheets in an ontology and record our approach in heuristics. We interview the original developers of the spreadsheets to compare our findings with their views. Our reconstructed conceptual model does not conflict with the developer's views, but represents a different perspective, as the developers are primarily focussed on the calculation workflow.

Keywords: ontology, conceptual model, spreadsheet, workflow, implicit knowledge

1 Introduction

In this paper we show that it is possible to reconstruct the conceptual model of a research project from its spreadsheet implementation.

Information technology enables scientists to collect, manipulate, and communicate ever increasing amounts of data. Scientists are encouraged to make their research data publicly available [1]. This results in an tremendous and rapidly growing amount of scientific data available on the web. It is hard to make efficient and effective use of these data, because the meaning and context of the data are often not clear to people that did not produce them. To enable integration and reusability of scientific data on the web, there is a need for semantic annotation of both the datasets and the associated research.

Significant progress in this direction has been made in the past years, especially in the field of bioinformatics. Formal descriptions of the scientific workflow support scientists to integrate and analyze data [2]. The W3C provenance working group[3] has developed a open provenance model, PROV, to document,

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

share and process provenance information, which helps scientists to ensure reproducibility of their analyses [3]. And several methods are available to link entities in datasets with concepts in the Linked Open Data Cloud [4,5].

Our focus in this study is on the semantic description of spreadsheets. Spreadsheets are one of the main tools used by scientists to store and analyze research data. A drawback of current spreadsheets is that their free format leads to sloppy or limited specification of the semantics of the data and calculations [6,7]. Scientists have the possibility to annotate their spreadsheets using tools like RDF123 [7], XLWrap [8] and "OM Excel add-in" [6]. These programs are primarily concerned with the values of the cells in the spreadsheet. It is not yet possible to formally describe the semantics of the conceptual model, i.e., the used concepts and their interrelations, or the meaning of concrete calculations in a spreadsheet. This information is sometimes available in documentation, such as articles, presentations, and technical reports, but is always present in the heads of the researchers. Semantic annotation with concepts from the underlying conceptual model could facilitate the reuse of data captured in scientific spreadsheets.

In this paper we explore to which extent the conceptual model of a research project can be recognized in its spreadsheet implementation and to which extent the reconstructed conceptual model agrees with the image inside the researcher's head. To study this we perform a case study in the domain of environmental science (section 3). We study Excel spreadsheets from an existing research project and try to manually reconstruct the underlying conceptual model. We record the steps we take during the reconstruction to investigate opportunities for semi-automated support of spreadsheet annotation (section 4). We analyze our results (section 5) and and verify our reconstructed conceptual model with the original developers of the spreadsheets (section 6). In sections 7 and 8 we evaluate the case study, summarize our findings and list remaining issues for future research

2 Related Work

Methods and tools to semi-automatically derive semantic descriptions from data have been developed for several scientific media, like RDF123 [7] and XLWrap [8] for spreadsheets, D2RQ [9] for databases and OntoLearn [10] and KAON[4] for text documents. Annotation of the scientific content of the data and methods has been done at various levels. Document-level content annotations are being made both manually and automatically. Examples are automatically connecting biomedical documents to terms from the Gene Ontology [11] and semi-automatic annotation of geo-spatial datasets with metadata provided by international guidelines from INSPIRE [12]. A higher level of abstraction that is being investigated is the annotation of scientific discourse and argumentation [13]. The integration of semantic descriptions with the actual documents, amongst which

[4] KAON, `http://kaon.semanticweb.org/`

spreadsheets, is facilitated by the Open Document Format 1.2,[5] which is based on RDF.

3 Case Study

Our case study is a scientific model for energy policy analysis [6], developed by the PBL Netherlands Environmental Assessment Agency and the Energy Research Centre of the Netherlands (ECN). We study spreadsheets from this model, try to manually reconstruct the underlying conceptual model model, and verify our results with the original researchers.

Procedure For our analysis we select two spreadsheets of different types. We adopt an open-minded view towards the content of the spreadsheets and consider the values, terms and formulas as well as their relative and absolute location in the sheet. We analyze the various patterns in the spreadsheet and determine to what extent these patterns provide insight in the semantics of the content. From the point of view of knowledge reuse we decide to make an instantiation/specialization of an existing ontology, the OM Ontology for units of Measure and related concepts [6] During the process we observe the consecutive steps needed to recognize the semantics and record them in heuristics. The semantics are visualized by a coloring of the OM concepts in the spreadsheet and relating them in diagrams. In an interview with the developers of the spreadsheets we compare their ideas with the content and configuration of our semantic model.

Data sources The case study model is developed to explore design options for the Dutch energy system in 2050 and calculate consequences in terms of greenhouse gas emissions and production costs. Model calculations as well as input and supporting data are represented in several interconnected Excel worksheets. Both spreadsheets and data are property of the authoring institutions and are not publicly available. In this study we analyze one spreadsheet from the calculation workbook, the *calculation sheet*, and one from a data-workbook, the *data sheet*, both on the subject 'traffic'.

Ontology We choose the OM ontology[7] [6] to formally describe the semantics of the analyzed spreadsheets. A pragmatic reason to select this ontology is that it is developed by close colleagues who are around to provide support in its application. OM as is an existing ontology designed for science and engineering practice. It contains the types of concepts that are relevant for the studied domain as it describes quantities, measures and units of measure as well as the natural phenomena they are related to. As such OM allows both statements about the structure of the physical world and corresponding quantitative observations, which is essential to construct a conceptual model of domain knowledge.

[5] ODF, `http://opendocumentformat.org/`
[6] Edesign, `http://www.pbl.nl/e-design/`
[7] OM ontology, `http://www.wurvoc.org/vocabularies/om-1.6/`

4 Ontology Reconstruction

Semantic Characterization We start the analysis of the spreadsheets with characterizing the terms in the spreadsheet as instances of OM concepts (Figure 1). We assume that each term has a unique definition an semantic characterization in the context of the model (Appendix rule 1). Four main concepts from the OM ontology are recognized in the spreadsheets: *Phenomenon*, *Quantity*, *Unit of Measure* and *Measure*. We add the concept *Quality* (Appendix rule 6), which describes a qualitative property of a *Phenomenon*. Part of the terms in the spreadsheets can easily be characterized as *Phenomenon* and *Quantity*. For example, in Figure 1, MicroEV, a technology used by cars, is a real world process and can therefore be characterized as *Phenomenon*. InvestmentCost is a property of MicroEV which can be observed and quantified and can therefore be characterized as *Quantity*. Because of their appearance the numbers and symbols or codes in the spreadsheets can be characterized as *Measures* and *Units of Measure* respectively (Appendix rule 2,3).

With semantics of part of the terms in the spreadsheet already known, the semantics of the remainder of the terms can be deduced applying multiple heuristics at the same time. Main assumption here is that *Measures* are always related to *Units of Measure* and *Quantities*, *Quantities*, again, are always related to *Phenomena* (Appendix rule 8,9,10). Also the design of the tables in the spreadsheets provides information about the semantics of the terms (Appendix rule 11 to 14).

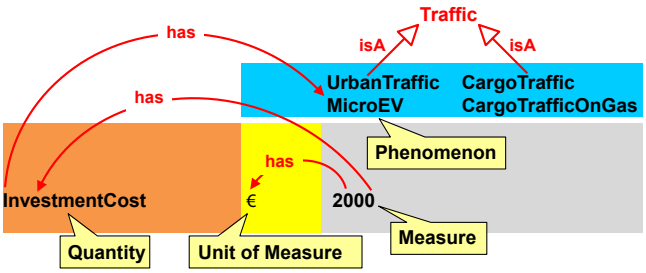

Fig. 1. Example, in outline, of the color markup of the main used concepts and their relations in one of the spreadsheet tables

Indication of Hierarchical and Property Relationships Next step is to define mutual relationships between the recognized concepts in the spreadsheets (Figure 1). The hierarchical relationships are mainly expressed in the design of the spreadsheet tables (Appendix rule 19 to 21).

As mentioned in the previous section from OM ontology we already have a basic assumption on the property relationships between the concepts in the spreadsheets. These relationships are also expressed in the design of the spreadsheet tables (Appendix rule 15 to 18).

Analysis of Formulas Final step is to analyze the formulas in the spreadsheet. These are not directly visible, but are 'hidden behind' the numbers in the spreadsheet cells. To understand the conceptual meaning of the formulas in the spreadsheet, we look at the *Quantities* that are related to the *Measures* in the formula, and to the *Phenomena* that are related to those *Quantities*. Analyzing the formulas reveals additional information on relationships between concepts an also helps deducing implicit information.

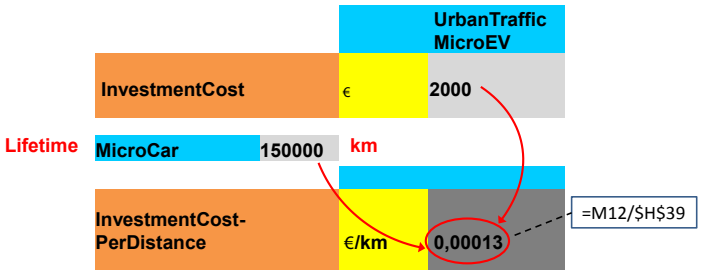

Fig. 2. Example of a formula connecting different concepts in the spreadsheet

An example is presented in Figure 2. The *Unit of Measure* and *Quantity* of the statement 'MicroCar 150000' are missing. From the other concepts in the formula we can deduce that the missing *Unit of Measure* is 'km'. From the related *Unit of Measure* 'km' and the related *Phenomenon* 'MicroCar' we can deduce that the missing *Quantity* is 'distance driven'(Appendix rule 4). The textual comments in the spreadsheet indicate that it concerns the distance driven during the lifetime of a car. Summarized in natural language, the formula shows that the investment costs per distance for MicroCars using MicroEV Technology are equal to the total investments costs divided by the total distance driven by a MicroCar during its lifetime. From the formula we also deduce that Microcars are a type of Urban Traffic, and that the car Technologies present in the spreadsheet can be categorized by the type of Traffic that is using them (Figure 3).

5 Characteristics of the analyzed spreadsheets

Results of our study are the following: 1) An OWL ontology based on OM[8] 2) A list of heuristics to extract semantic information from the spreadsheets 3) A visualization of the main OM concepts by color markup of the spreadsheets[9], 4) A visualization of the conceptual model of the case study (Figure 3). 5) A visualization of the workflow in the spreadsheets[10]

Additionally, we performed a rough quantitative analysis on our findings and examined the workflow in the spreadsheets.

[8] Edesign ontology, `http://semanticweb.cs.vu.nl/edesign/`
[9] Spreadsheet Color Markup, `http://semanticweb.cs.vu.nl/edesign/`
[10] Workflow, `http://semanticweb.cs.vu.nl/edesign/`

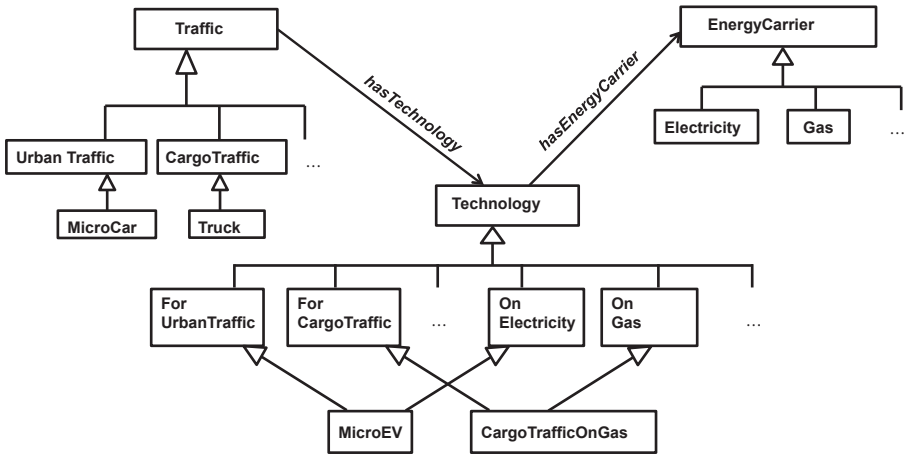

Fig. 3. Overview of the main classes, subclasses and relations of the *Phenomena* used in the analyzed spreadsheets.

Concepts The *Phenomena* recognized in the spreadsheets can be categorized into three main classes 'Traffic', 'EnergyCarrier' and 'Technology' (Figure 3). All *Phenomenon* classes that are used in the *calculation sheet* are also used in the *data sheet* (Table 3), but in the *data sheet* more subclasses are distinguished (Table 1). Main explanation is that the *data sheet* includes all possible traffic technologies, while the *calculation sheet* focuses on passenger traffic only. Both sheets use about the same number of *Quantity* classes (Table 1) but only 7 of these are used in both sheets (Table 3). The majority of the *Quantity* classes in both sheets are properties of the *Phenomenon* 'Technology'.

Table 1. Number of classes per type represented and used in the analyzed Excel sheets

Type	Data sheet			Calculation sheet		
	Total	Representation		Total	Representation	
		Explicit[a]	*Implicit*[b]		*Explicit*[a]	*Implicit*[b]
Phenomenon	33	24	9	20	14	6
Quantity	27	6	21	30	19	11
Quality	1		1	3	2	1

[a] explicitly mentioned in Excel sheet
[b] deduced according to heuristics in appendics

Workflow We distinguish between different formula types by looking at the *Quantities* they use as in and output. The *calculation sheet* uses more types of formulas than the *data sheet*(Table 2; the formula types in both sheets are different as the sheets play different roles in the workflow of the entire model. The majority of the formulas in both sheets calculate new properties of *Phenomenon* 'Technology', for which they may use other properties of *Phenomenon* 'Technol-

ogy' as well as properties of other concepts. The formulas that are represented and used in the *calculation sheet* are all connected with each other, while in the *data sheet* only 2 formulas are connected.

By analyzing the connections between the formulas we could determine the final output *Quantities* of the sheets, i.e.,production costs, emission of greenhouse gases and required energy input, and how they were derived. In essence we could deduce the workflow[11]in the spreadsheets .

Table 2. Represented and used formulas in the analyzed Excel sheets

	Data sheet	Calculation sheet
nr. formulas		
Total	4	10
Output related to Technology[a]	4	7
Input related only to Technology[b]	2	3
Input related to Technology and other concepts[c]	2	3
nr connections between formulas	1	11

[a] Formulas with an output *Quantity* related to *Phenomenon Technology*
[b] Formulas with only input *Quantities* related to *Phenomenon Technology*
[c] Formulas with both input *Quantities* related to *Phenomenon Technology* and input *Quantities* related to other *Phenomena*

Table 3. Similarities in used entities between the analyzed Excel sheets

Entity	Phenomena	Quantities	Qualities	Formulas
nr used in both sheets	20	7	1	0

6 Interview with Model Developers

We interviewed the developers of the spreadsheets to compare their view on the conceptual model with the content and configuration of our semantic model.

Workflow The researchers mentioned the calculation of greenhouse gas emissions and production costs as the main aims of their model. The calculation procedure in the model is based on assessing supply and demand of EnergyCarriers. In the model structure they did not make a distinction between these two aspects as they assume that Technologies and Sectors could contribute to both supply and demand of EnergyCarriers. However, they did use a clear distinction between supply and demand in the model workflow

We could derive the aims of the analyzed sheets from the final output *Quantities* of the formulas and these matched with the aims the model developers mentioned for the entire model. We did not notice the emphasis on the supply and demand equilibrium in the workflow, nor a distinction between supply and demand elements in the model. This may be caused by our limited scope as

[11] Workflow, `http://semanticweb.cs.vu.nl/edesign/`

we focused on one energy demanding sector, 'Traffic', and did not consider the workflow in the model as a whole, which consisted of many more spreadsheets.

Concepts According to the developers the main concepts in the model are the EnergyCarriers and the Technologies that use these EnergyCarriers as in or output. The developers designed a template for the spreadsheets in the model. It is easy to recognize semantics of the terms in the spreadsheet as they all have known, fixed locations. The model is perceived by the developers mainly from the point of view of the calculation workflow.

In our analysis we also distinguished EnergyCarriers and Technologies as important concepts and recognized their association. We included Traffic, one of the energy-demanding sectors, as third main concept. This did not match the view of the developers, which, again, may be caused by the limited scope of this study. During the reconstruction process we noticed the influence of the location of the terms in the spreadsheet on their semantics and already recorded this in our heuristics. We modeled primarily from the perspective of the concepts and their relations, as opposed to the calculation workflow. Due to this difference in point of view, it was difficult to verify the reconstructed conceptual model. However, we did not recognize any inconsistencies between our interpretation of the conceptual model and that of the developers.

7 Discussion

Ontology reconstruction The design of the spreadsheet contained implicit but valuable information on the underlying conceptual model. The heuristics we formulated appeared successful in extracting this information from the two analyzed spreadsheets. We think that the analysis of the formulas offers several opportunities for semi-automated support of spreadsheet annotation. Formulas are easily recognizable and they provide conceptual knowledge as they are connected to concepts in the conceptual model. Furthermore, analysis of formulas, as part of the calculation workflow, matches the approach of most model developers. Analysis of the formulas in tabular data is, however, only possible when they are stored in Excel or ODF format; the CSV format which is often used in research on spreadsheet semantics cannot store formulas.

Our methodology might be applicable to spreadsheets from other domains, if they contain empirical data with corresponding units of measure and especially if they are also used for data analysis, i.e., when they contain formulas. Testing our methodoloy outside the scope of this study could be a subject for further research.

Meta ontology OM proved to be a suitable ontology to formally describe the semantics of the analyzed spreadsheets. We suggested a few additions to OM, especially on the level of concept properties, to facilitate semantic characterization of terms(Appendix rule 6 to 10). A drawback of OM is that it is not yet grounded in foundational ontologies [14] A suitable alternative would have

been the QUDT ontology [12], which is comparable to OM but has more extensive foundation. QUDT does not include phenomena which we think are essential to formally describe conceptual scientific knowledge. The developers of OM note possibilities for integration of their ontology with QUDT [14] so we would have needed to specialize QUDT in order to be able to use it.

Research questions We were able to reconstruct a conceptual model, i.e., the used concepts and their relations, from the analyzed spreadsheets. We could not find inconsistencies between our reconstructed conceptual model and the story of the developers of the spreadsheets.

From the interview it was clear that the developers were primarily focused on the calculation workflow, and showed limited interest in the conceptual model. Although we set out to reconstruct the conceptual model in the developer's head, we found that we were constructing a model that did not exist yet. This is in line with observations made by Clancey[15]. The differences in focus may be explained by different viewpoints on the purpose of these types of scientific models. The model developers may see these models mainly as instruments to perform scientific analyses, while we consider them to be tools to communicate scientific knowledge.

8 Conclusions and Future Work

With the case study described in this paper we show that it is possible to reconstruct the conceptual model of a research project from its spreadsheet implementation. The formulas in the spreadsheets contain implicit knowledge about both its semantics and the calculation workflow. The developers were focussed on this workflow rather than on the conceptual model of their research. For them an explication of semantics is not strictly needed, but it was important for us to check wether the addition led to conflicts.

In future work we intend to investigate to what extent reconstructing conceptual knowledge captured in scientific models is helpful in understanding and reusing these models. Visualization of the reconstructed conceptual model would in that case be a feasible next step. We also plan to explore which parts of the scientific modeling workflow could benefit from semantic annotation and how this can be facilitated trough semi automated support. Linking of concepts to concepts used in the Linked Open Data cloud and formalization of the data in schema's, such as the RDF Data Cube vocabulary,[13] OM, and QUDT, could facilitate the reuse of both concepts and data from this case study. It could support quick and easy connections between *data sheets* and *calculation sheets* within the project and enable connections with other research projects elsewhere in the world. The actual porting of the data is beyond the scope of this paper and future work. We think, however, that the limited interest of scientists in the conceptual model of their research project could prove to be a bottleneck

[12] QUDT, `http://www.qudt.org`
[13] Data Cube, `http://www.w3.org/TR/vocab-data-cube/`

for reuse of scientific data. If and how this is the case is also a subject of future work.

9 Appendix

1. Each term has a unique definition and semantic characterization in the context of the model.
2. *Units of Measure* is represented as symbol/code
3. *Measure* is represented as numerical value
4. The energy domain has commonly used combinations of (*Phenomenon-*)*Quantity-Unit of Measure*; with one concept present, the others can be deduced

These rules describe proposed additions to OM ontology:

6. *Quality* is a qualitative property of a *Phenomenon* which can be observed
7. *Quality* has a necessary, functional property 'phenomenon' with range *Phenomenon*
8. *Measure* has a necessary, functional property 'quantity' with range *Quantity*
9. *Measure* usually has a functional property 'unit of measure' with range *Unit of Measure*
10. *Quantity* has a necessary, functional property 'phenomenon' with range *Phenomenon*

These rules are on design of the spreadsheet (tables):

11. The body of an Excel table can only have *Measures*
12. Instances of *Quantity*, *Quality* and *Phenomenon* are only present in header rows or columns.
13. If one of the terms in a list can be semantically characterized as a certain concept, the other terms can be characterized as the same concept
14. If in a list in a header column or row only one term is present, it is valid for the whole list.
15. A *Quantity* related to a *Measure* is the nearest represented in the same row or column
16. *Measures* in table body are related to the terms in the headers of the same column and row
17. The *Phenomenon* related to a *Quantity* is the nearest one represented in the same column or row (as the *Measure* that that *Quantity* is related to)
18. If some *Quantities* in a header row or column are related to a *Phenomenon*, the other *Quantities* in that row or column are related to the same *Phenomenon*
19. *Phenomena* that are in the same (horizontal or vertical) list are subclasses of one parent class
20. A *Phenomenon* that is not related to a *Quantity*, but is represented in the same row or column as another *Phenomenon*, is a parent class of that *Phenomenon*.
21. A *Quantity* that is not related to a *Measure* or *Phenomenon*, but is represented in the same row or column as another *Quantity*, is a parent class of that *Quantity*

References

1. Boulton, G., Rawlins, M., Vallance, P., Walport, M.: Science as a public enterprise: the case for open data. Lancet **377**(9778) (May 2011) 1633–5
2. 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
3. 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
4. Ngonga Ngomo, A.c., Auer, S.: Limes-a time-efficient approach for large-scale link discovery on the web of data. Proceedings of IJCAI (2011) 2312–2317
5. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk A Link Discovery Framework for the Web of Data. (2009)
6. Rijgersberg, H., Wigham, M., Top, J.: How semantics can improve engineering processes: A case of units of measure and quantities. Advanced Engineering Informatics **25**(2) (April 2011) 276–287
7. Han, L., Finin, T., Parr, C., Sachs, J., Joshi, A.: RDF123 : From Spreadsheets to RDF. (2008) 451–466
8. Langegger, A., Wolfram, W.: XLWrap Querying and Integrating Arbitrary Spreadsheets with SPARQL. (2009) 359–374

9. Bizer, C., Seaborne, A.: D2RQ Treating Non-RDF Databases as Virtual RDF Graphs. In: Proceedings of the 3rd International Semantic Web Conference (ISWC2004). (2004)
10. 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)
11. Smith, T.C., Cleary, J.G.: Automatically linking MEDLINE abstracts to the Gene Ontology. In: Proc. ISMB 2003 BioLINK Text Data Mining SIG. (2003) 1–4
12. Macário, C.G.N., de Sousa, S.R., Medeiros, C.B.: Annotating geospatial data based on its semantics. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09, New York, New York, USA, ACM Press (2009) 81
13. Shum, S.B., Clark, T.: Scientific Discourse on the Semantic Web : A Survey of Models and Enabling Technologies. Semantic Web Journal (2010)
14. Rijgersberg, H., Van Assem, M., Top, J.: Ontology of units of measure and related concepts. Semantic Web **0** (2012) 1–11
15. Clancey, W.J.: The epistemology of a rule-based expert system a framework for explanation. Artificial Intelligence **20**(3) (May 1983) 215–251