Conceptual Modelling and Ontologies for Biology: experiences with the bacteriocin database

C. Maria Keet
School of Computing, Napier University, 10 Colinton Road, EH10 5DT, Scotland,
+44 131 455 2773, m.keet@napier.ac.uk

Abstract

This article outlines some characteristics of biological data, which affect its modelling as well as ontologies in the subject domain of biology (including ecology and agriculture), which in turn influence the quality of developed software and the reusability of the models. These aspects arose from a literature review and a case study of developing a bacteriocin database. Conflicting goals in software requirements and aspects for future research are highlighted.

Background

Data characteristics

In an abstract sense, one can consider ‘data is data’ and its principles are the same regardless the subject domain. However, there are distinct features of biological data influencing conceptual modelling, ontology development and ontology integration that are less important, or absent, from the examples found in the research literature, often addressing ‘common sense’ domains where the modeller is also subject matter expert, such as the modelling and integration of ontologies of universities. Refer to Keet (2003b) for discussion on these five general factors characteristic to biological data. In addition, several factors are identified that are more prevalent within the ecological and agricultural domains: ecology and agriculture can comprise interdisciplinary (Mode 2) science with ‘interfering’ management and policy perspectives, how one would want or should see the world, and the broader social views of the natural environment (Argent, 2003 *in press*; Gangemi *et al*, 2002). There is an ‘embeddedness’ of mathematical formulas within ecological and agricultural concepts and their use which draws multiple concepts together; e.g. the canopy photosynthesis (Keller and Dungan, 1999) or Monod kinetics for organism growth under nutrient limitation. Further, uncertainties exist in many intertwined system parameters that still need to be included in the model and the system has to be able to cope with unavailable information by using estimates (Huang and Chang, 2003). Last, note that for each specific subject domain there are additional challenges to be resolved, e.g. classification systems in plant taxonomy (Raguenaud *et al*, 2002; Priss, 2003) or the loosely defined groups of microorganisms (Keet, 2003a).

Modelling paradigms in Informatics

Apart from biological data characteristics and domain heterogeneity (Keet, 2004), there are multiple modelling paradigms for semantically, structurally and syntactically representing concepts and their relationships. Here it is important to note the difference between ecological modelling and the [progress in] modelling paradigms within the discipline of informatics. Within agricultural and ecological modelling, there is a plethora of methodologies and graphical representations, e.g. Odum’s conventions, that do not bear any relation to informatics models and are more focussed on ecology and simulations than modelling for its own sake. The software applications have been gradually shifting from procedural legacy systems to Object-

Oriented software[1] and relational databases, followed by the recent development of an Ecological Metadata Language[2] and ontologies, e.g. AOS[3] and SEEK[4], to annotate and model ecological and agricultural knowledge. Ontologies are of special importance, because they capture the semantics in an implementation independent way, its generalization of knowledge enabling a higher degree of reuse. Reusing extant knowledge captured in an ontology, of e.g. a simulation of nutrient recycling, for some other software application does not require re-analysing and re-modelling the subject domain, but building further on the foundations laid by previous research, hence moving forward more swiftly. Informatics modelling paradigms such as Object-Orientated or Entity-Relationship are closer to computational models than 'true' conceptual modelling approaches such as Object Role Modelling (e.g. Halpin, 2001), Formal Concept Analysis and Conceptual Graphs (Juristo and Moreno, 2000). However, even the latter tend to be used with the particular application in mind. In contrast, ontologies represent what is or what can be, regardless how this information is to be used, and capture *consensus* between subject matter experts.

How conceptual modelling, ontologies and modelling biological data can translate to 'reality' was investigated with a case study of developing the bacteriocin database.

Materials and Methods

A biological database was developed via an iterative process, on request for Dr. Scannell from the Department of Food Science[5]. This bacteriocin database was the first attempt of its kind to represent these biological semantics in a conceptual model within the application of food microbiology. Her main requirements were to have an easily accessible, structured and searchable repository for bacteriocin-related data extracted from the different kinds of journal articles (food safety, genetics, microbiology and so forth), such as the bacteriocin-producing bacteria and their relevant genes, food products and mode(s) of action of bacteriocins. The primary data analysis and modelling technique was ER, augmented with ORM. The database was implemented with DBMS InfoMaker (refer to Keet (2003c) for details). This was augmented with literature research on ontologies and conceptual modelling techniques.

Results and Discussion

The database was successfully developed from the perspective of the customer, hence achieved its goal (Keet, 2003c). However, from a computing perspective, several competing goals did interfere. The understandability of the ER model by the domain expert: it is known (e.g. Aguado *et al*, 1998) that domain experts do not know how to formalise their knowledge well and to address this, one can teach this to the domain experts, or the computing scientist knows/learns enough of the UoD (like with this case study), or one can develop an intermediate representation to meet half-way. The latter approach was taken by deploying ORM and using the near-natural language included with its modelling tool VisioModeler. Either way, this is a significant practical hurdle for developing biological applications. Secondly, the

[1] Consult e.g. Mineter *et al* (2003), Baskent *et al* (2001) and the Analest and Reciclado de Nutrientes of the ICA (http://www.ica.inf.cu).

[2] By the Knowledge Network for Biocomplexity: http://knb.ecoinformatics.org/software/eml/.

[3] Agricultural Ontology Service http://www.fao.org/agris/aos/.

[4] Science Environment for Ecological Knowledge, http://seek.ecoinformatics.org.

[5] Faculty of Agriculture, University College Dublin, Ireland.

subject domain is an *applied science*: capturing the subject domain semantics of an applied bioscience faces different problems compared to conceptual modelling for the 'core' life sciences, because the former requires an emphasis on practical solutions conceptually representing the integration of various fields, whereas the latter stresses conceptual and ontological 'all-inclusive' models within their primary specialisations like biochemistry and genetics (Keet, 2003a). For this reason, abstractions as represented in the GOC[6] or AOS ontologies were explored but not used, because the integrative domain of food science cannot simply use a combination of ontologies to construct a conceptual model: the food groups would be of a *descriptive* kind, whereas a bacteriocin with its mode of action is of a certain *function* with *activities*. Deploying both types of ontologies in orthogonal manner would create excessive classifications, which are deemed not relevant by the customer and would have to be pruned manually. Similar problems exist within other sections of agriculture. In addition, if one were to use ontologies for constructing a conceptual model, this would require preceding *integration* of ontologies, which is a major problem area in itself because the (semantically) different views still exist even on the ontological level. Recently published research by Jarrar *et al* (2003) might alleviate this and facilitate more reuse of knowledge whilst catering for diverging semantics, by distinguishing between the *ontology base*, recording concepts and their relationships, and *ontological commitments*, defining the rules how concepts are used[7], thereby providing a link to conceptual modelling.

Due to the location, size and subject matter of the bacteriocin project, no integration of conceptual models (or even ontologies) was required; although it may be advantageous if a repository of models existed, where one simply could select relevant *sections* of a larger body of reusable knowledge to create a conceptual model. Notwithstanding, if one assesses biological data, and agricultural data in particular, it is highly 'localised' where different perceptions and particularities of ecological and agricultural data cannot be represented as the same thing across natural language barriers: similar structural representations may involve semantic conflicts on closer inspection (Keet, 2004).

From this case study, one might think ontologies are certainly not a panacea, but a higher level of abstraction and reusability were not the primary aim – though as a *researcher* this would be more interesting. The trade-off between the competing goals when developing the bacteriocin database was in favour of developing a working database to the client's satisfaction within a limited time frame. Nevertheless, the present conceptual models (ER and ORM) still allow for relatively straightforward subsequent bottom-up ontology development so that other scientists may take advantage of the existing modelled domain knowledge in the near future.

Future research

Although the bacteriocin database was successfully implemented, the process uncovered several suggestions for further research. Some of these aspects will be pursued in the near future: generalising biological knowledge taken from conceptual models for the bottom-up development of ontologies and integration with existing generic ontologies and, possibly, the effect of natural language differences in conceptual modelling and ontology development.

[6] Gene Ontology Consortium, http://www.geneontology.org.

[7] Refer to Keet (2004) for an example of this approach, addressing a section of microbiology.

References

Aguado, G., Bañón, A., Bateman, J., Bernardos, S., Fernández, M., Gómez-Pérez, A., Nieto, E., Olalla, A., Plaza, R. and Sánchez, A. (1998). ONTOGENERATION: Reusing domain and linguistic ontologies for Spanish text generation. *Proceedings of the ECAI'98 Workshop on Applications of Ontologies and Problem Solving Methods*, Brighton, U.K.

Argent, R. M. (2003 in press). “An overview of model integration for environmental applications—components, frameworks and semantics.” *Environmental Modelling & Software*, **xx**(xx): xx-xx.

Baskent, E.Z., Wightman, R.A., Jordan, G.A., Zhai, Younghua (2001). “Object-oriented abstraction of contemporary forest management design.” *Ecological Modelling*, **143**: 147-164.

Gangemi, A., Fisseha, F., Pettman, I., Pisanelli, D.M., Taconet, M., Keizer, J. (2002). A Formal Ontological Framework for Semantic Interoperability in the Fishery Domain. *Proceedings of the ECAI-02 Workshop on Ontologies and Semantic Interoperability*, Lyon, France.

Halpin, T. (2001). *Information Modeling and Relational Databases*. San Francisco: Morgan Kaufmann Publishers.

Huang, G. H. and Chang, N.B. (2003). “Perspectives of environmental informatics and systems analysis.” *Journal of Environmental Informatics*, **1**(1): 1-6.

Jarrar, M., Demy, J. and Meersman, R. (2003). “On Using Conceptual Data Modeling for Ontology Engineering.” *Journal on Data Semantics Special issue on "Best papers from the ER/ODBASE/COOPIS 2002 Conferences"*, **1**(1): 185-207.

Juristo, N., and Moreno, A.M. (2000). “Introductory paper: Reflections on Conceptual Modelling.” *Data & Knowledge Engineering*, **33**(2): 103-117.

Keet, C.M. (2003a). *Conceptual Modelling for Applied Bioscience: The Bacteriocin Database*. CSPS Nr 0310001, 8 October 2003. http://www.compscipreprints.com/comp/Preprint/mkeet/20031008/1.

Keet, C.M. (2003b). “Biological data and conceptual modelling methods”. *Journal of Conceptual Modeling*, **29**, October 2003. http://www.inconcept.com/jcm.

Keet, C.M. (2003c). *The use of bacteria and bacteriocins in the food industry – modelled and documented in a relational database*. BSc Final Year Project, Open University, UK.

Keet, C.M. (2004). *Aspects of ontology integration*. School of Computing, Napier University. January 2004. http://www.dcs.napier.ac.uk/~cs203/AspectsOntoIntegration.pdf.

Keller, R.M. and Dungan, J.L. (1999). “Meta-modeling: a knowledge-based approach to facilitating process model construction and reuse.” *Ecological Modelling*, **119**: 89-116.

Mineter, M.J., Jarvis, C.H. and Dowers, S. (2003). “From stand-alone programs towards grid-aware services and components: a case study in agricultural modelling with interpolated climate data.” *Environmental Modelling & Software*, **18**(4): 379-391.

Priss, U., (2003), ‘Formalizing Botanical Taxonomies’. *Proceedings of the 11th International Conference on Conceptual Structures, 2003*. Online preprint: http://www.upriss.org.uk/papers/iccs03.pdf. Date accessed: 30-9-2003.

Raguenaud, C., Pullan, M.R., Watson, M.F., Kennedy, J.B., Newman, M.F. and Barclay, P.J., (2002), “Implementation of the Prometheus Taxonomic Model: a comparison of database models and query languages and an introduction to the Prometheus Object-Oriented Model”. *Taxon*, **51**: 131-142.