Semantic Integration of Environmental Models for Application to Global Information Systems and Decision-Making

D. Scott Mackay
Department of Forest Ecology & Management, and
Institute for Environmental Studies
University of Wisconsin - Madison
1630 Linden Dr. Rm. 120, Madison, WI 53706
dsmackay@facstaff.wisc.edu

Abstract

Global information systems have the potential of providing decision makers with timely spatial information about earth systems. This information will come from diverse sources, including field monitoring, remotely sensed imagery, and environmental models. Of the three the latter has the greatest potential of providing regional and global scale information on the behavior of environmental systems, which may be vital for setting multi-governmental policy and for making decisions that are critical to quality of life. However, environmental models have limited prootocol for quality control and standardization. They tend to have weak or poorly defined semantics and so their output is often difficult to interpret outside a very limited range of applications for which they are designed. This paper considers this issue with respect to spatially distributed environmental models. A method of measuring the semantic proximity between components of large, integrated models is presented, along with an example illustrating its application. It is concluded that many of the issues associated with weak model semantics can be resolved with the addition of self-evaluating logic and context-based tools that present the semantic weaknesses to the end-user.

1 Introduction

A major challenge facing the human race is global change and its affects on quality and diversity of life. Global information systems have the potential of providing timely, integrated regional to global scale information needed by policy and decision makers on environmental issues. For example, the global information infrastructure (GII), which is based on inter-networking technology and the world-wide web is currently demonstrating that global information systems will benefit problems that require a synthesis of information from diverse sources if end-users have the ability to tie the information together in a meaningful way. With the progress in global inter-connectivity, we now need to deal with more heterogeneous information consisting not only of a broader variety of digital data, but also operations, such as simulation models, which create new data and information. The scale of the problem has changed from a few databases to millions of information resources, and the new resources are added independently to the accessible set of resources, as other resources change rapidly or disappear.

Geographic information systems (GIS) and simulation modeling, both non-traditional information systems applications, can deliver spatially and temporally diverse data and information to the global community. Given our limited understanding of complex earth systems and limited resources with which to collect and store data about the earth, it is conceivable that a major component of future global information systems will support environmental policy through direct or indirect use of environmental models and their application at a variety of spatial scales. These applications pose significant integration challenges since differences in spatial and temporal scales of data, information, and models are difficult to reconcile. Current integrated environmental models present are analogues for these future, heterogeneous information system, as they are typically constructed in a bottom-up fashion, using existing simpler models that describe only a small part of the earth system. This current approach to building a knowledge base of earth system processes is not unlike what global information systems hope to provide; it brings together the best and brightest results from multiple earth science disciplines. However, current, and hence future, model integration must reconcile semantic differences between models if their results are to be interpreted.

Given the argument that current integrated environmental models typify the sort of heterogeneous simulation information sources that global information systems promise, this paper addresses the issue of semantic integration in the context of current environmental models. It addresses concerns about spatial and temporal conflicts between environmental models, the weakly defined semantic connections between real world earth systems and their model counterparts, and the reasoning processes needed to resolve semantic conflicts as integrated models developed for a limited range of problems are applied more broadly. Ultimately, global information system must integrate models developed for the purpose of addressing a variety of scales. A query-directed semantic integration for a specific integrated environmental model is highlighted in order to illustrate the idea of model self-evaluation and semantic proximity as tools for measuring and reconciling semantic conflicts and for presenting these results in manner that is easily understood by model end-users. The approach taken here assumes that sub-models within an integrated system have a measurable semantic proximity, which can be associated with each query result. Semantic proximity is provided as feedback from the integrated models. This feedback in turn is used to determine (1) acceptance or rejection of the integrated model for application to a specific query, and (2) identify a need for new models for which semantic conflicts can be resolved. The next section lays some fundamental ideas on how ontology, semantics, and context are applied in the context of environmental models. This is followed by a discussion of the major issues facing model information management, and then a specific example.

2 Ontological And Semantic Basis

An ontology can be thought of as a way in which an agent views the world, the features in the world and the processes that govern the dynamics of those features. This definition parallels that of Lee and Siegel (1996) who provide their interpretation of Bunge's (1979) theory of ontology. An ontology may consist of complex dynamic systems, which may be decomposed into simpler systems and ultimately physical things that have measurable properties. Only measurable properties are considered part of an ontology. That which can be measured depends upon the spatial and temporal scale (or likewise the respective grain) of interest to an agent, suggesting that different agents viewing the same environmental system may each have a different ontology. For instance, a field forest ecologist may view the forest canopy in terms of detailed properties of individual tree leaves. This detail is not observable from the perspective of a remotely sensed image analyst who sees only an aggregation of light reflected from the many leaves that comprise a vegetation canopy. Similarly, a modeler who is interested in the dynamic behavior of a stand of trees will "see" many details that are invisible to a regional or global scale modeler working at highly aggregated grid cells. These distinctive views that arise due to spatial aggregation apply also to the temporal domain, in which observations of dynamic behavior are normally limited by instrumentation precision and interests of the observer. The ability to measure properties about physical systems imposes restrictions that are reflected in model design.

We further distinguish between an ontology upon which individual information sources are constructed and the ontology of an end-user of the information sources. An end-user might interact with a collection of information sources by issuing a query that requires an integration of models, data, and analytical results. For example, the end-user may request information that is derived both from field-plot studies and from regional analysis of remotely sensed imagery. This is quite a common practice, as most models of complex environmental systems synthesize information sources from a variety of spatial and temporal scales. Furthermore, an end-user query might involve the synthesis of numerous information sources in the form of multiple models. This, too, is quite common. Regional and global simulation models that integrate plant physiological models and hydrological models in order to capture land surface - atmosphere exchange of carbon dioxide and water vapor must draw together models that were developed for very different purposes (and ontologies), and most likely different spatial and temporal scales. The interoperability of these models will depend to a large extent upon whether their respective semantics can be integrated or made cooperative in the context of an end-user query.

Semantics broadly refers to the system of represented objects and real world features they stand for, and ultimately how the objects and features are related. Relationships between objects in a representation and features in the real world are formed through the use of predicates, *e.g.*, forest stand or water flux, propositions, *e.g.*, a particular forest stands or measured water flux, and arguments, *e.g.*, vegetation type, volumetric unit of measurement, time of measurement, and scale of measurement. Meaning relations (Kashyap and Sheth,

1996) form a critical bond between objects; they may be integrity constraints, rules, programs, *etc*. In environmental models meaning is often carried in the explicit form of programs and the implicit underlying assumptions made by the model developer. It is differences in underlying assumptions that must be resolved when several application specific environmental models are combined to address larger problems.

The strength in meaning relations or relationships lies in their ability to portray the context in which a particular system semantics should be viewed (Kashyap and Sheth, 1996). It provides the real world semantics needed by the end-user who is trying to interpret the results obtained from the integrated information system. Context consists of a restricted set of propositions allowed, a set of predicates, and the domain or union of all reference classes of all predicates. To a modeler context is provided by state variables that describe stores and equations that describe fluxes, and by assumptions of spatial aggregation and time intervals that may be data-dependent or application-dependent. These assumptions usually provide important context information for the model developer, but are often not explicitly defined and so they are lost when the model is passed on to another user. Without proper context a model is easily misused. For instance, by providing inappropriate input data or, more innocuously, by combining the model with another model also with poorly defined context a semantic integration problem results.

3 Semantic Integration of Models

Environmental models, and information systems derived from these models, cannot be subjected to the same level of formal proof that is fundamental to traditional information systems applications (Oreskes *et al.*, 1994). Models, as abstractions of real world systems make simplifying assumptions, use heuristics, or simply make "leaps of faith" that enable them to capture the known properties of the real system while ignoring unknown. The presence of such incomplete or uncertain information embedded within each model that is then incorporated into a larger integrated system of earth systems as a whole, results in a semantic interoperability problem (Mackay *et al.*, 1996).

A growing interest in combining environmental models with GIS has resulted in the development of a number of prototype systems to address spatial model-base management issues. For instance, decision-support systems have aided in the application of environmental models (*e.g.*, Dunn *et al.*, 1996; Jamieson and Fedra, 1996). These decision-support systems support the selection, from a repository of models, those models best suited to a given problem. The selection process requires considerable human guidance, and so model description languages or interfaces have emerged to formalize some of the human decision processes (*e.g.*, Zeigler, 1990; Keller *et al.*, 1994; Mackay *et al.*, 1994; Bennett, 1997; van Deursen, 1995; Wesseling *et al.*, 1996). These authors describe intra- and inter-model linkages, constraints on inputs and outputs (*e.g.*, pre- and post-conditions), and system organization. However, they fall short of providing answers to semantic integration issues since they view the models as black boxes, as being constructed in a top-down fashion, or as trivial and non-representative of the type of environmental models that would be applied to global environmental issues. We have to consider the possibility of bottom-up design using complex models in a global information context. Ultimately, global information systems must rely on model and end-user autonomy, which is challenging in a spatial context (Worboys and Deen, 1991) and difficult to adapt in evolving environments (Ventrone and Heiler, 1991). The issues are considered in the context of a particular environmental model in the next section.

4 Semantic Proximity in Simulation

The Regional HydroEcological Simulation Systems, Dynamic (RHESSysD) is a spatial information processing and dynamic simulation toolkit for regional scale environmental modeling. It consists of a number of numerical models linked within a GIS framework, included a spatially aggregated distributed hydrological model that represents horizontal flow of water, and a spatially more detailed ecosystem model that accounts for vertical movement of water (Mackay and Band, 1997). The model is not unlike many integrated, regional environmental models, and so it is used here as a laboratory for exploring semantic integration problems associated with different spatial aggregation the sub-models.

We have previously described the use of model self-evaluation to identify semantic weaknesses, and a set of linguistic terms with which to describe how the semantic problems propagate to effect the end-user requested output from the model (Mackay and Robinson, 1998, 1999). The idea upon which this approach is based is that differences in spatial scale assumptions embedded within sub-models of an integrated whole lead to differences in model output interpretation. If these differences can be detected and quantified in a form that is amenable to portrayal in the form of a map, then the decision on whether to use the integrated model can be based on a sound understanding of its inherent semantic weaknesses at specific spatial locations. The processes involved are summarized in the flow diagram in Figure 1. The two most important processes involve using the Application Context, and the User Context Filters.

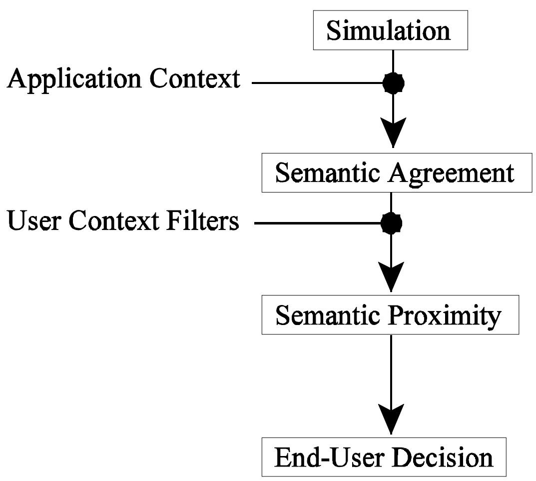

Figure 1. Process flow diagram showing the steps required in going from simulation to an end-user decision. The key processes are the Application Context and User Context Filters.

The Application Context establishes ground rules for the integrated model, including the input data and parameter space, and any rules or procedures describing the physical constraints that govern the modeled reality. As an example, we might state that water must flow from higher elevation points to lower elevation points within the bounds of a particular spatial domain. This constraint would not be in conflict with any user's view as long as the spatial domain is well defined. Mackay and Robinson (1999) present such a constraint in the form shown in Figure 2. A measurable semantic agreement in the water table position predicted by two spatially different sub-models of RHESSysD, is computed as the sum, ϵ_ϕ, of results of both rules applied to all simulated patches. However, in applications of this integrated model, semantic agreement between the different views of the simulated water table may not be of interest to an end-user of the system, and so this semantic agreement must be propagated to the variables of interest.

Rule 1 (Incorrect redistribution from patch i):

$$\forall\, j \in sub\text{-}area \mid elevation(j) > elevation(i),\quad i \in sub\text{-}area \;\Rightarrow\; \epsilon_{i_{upslope}} = \text{MAX}\left(0,\ \sum_{j=1}^{k} \Delta R_{j_i} w_j\right)$$

Rule 2 (Incorrect redistribution to patch i):

$$\forall\, j \in sub\text{-}area \mid elevation(j) < elevation(i),\quad i \in sub\text{-}area \;\Rightarrow\; \epsilon_{i_{downslope}} = \text{MIN}\left(0,\ \sum_{j=1}^{k} \Delta R_{j_i} w_j\right)$$

Figure 2. Example rules used for self-evaluation within RHESSysD Application Context. These rules calculate water table depth semantic agreement between an aggregated hydrological sub-model and spatially detailed stand-level vegetation water use sub-model. The term ΔR refers to a difference in recharge predicted by the two sub-models, and w is an areal weighting term.

Mackay and Robinson (1998, 1999) suggest that context for determining semantic agreement between sub-models be provided by a query, Q:

$$Q = \langle \alpha, \theta, \tau, \mu(\epsilon_\alpha) \rangle$$

where α is the goal of the query, θ and τ respectively define the spatial and temporal domains within which α is defined, and $\mu(\epsilon_\alpha)$ is a membership function that describes the semantic proximity of α as determined by a collection of sub-models. Proximity is presented in linguistic terms that are clearly understood by the model end-user. For instance, the term *semantically close* can be described using a fuzzy membership mapped over the range [0.0, 1.0]. α denotes a partial user context, but is only complete once $\mu(\epsilon_\alpha)$ is derived with User Context Filters.

User Context Filters consist of a one or more functions and associated linguistic terms that describe qualitatively how a context variable (α) responds to semantic agreement in a variable (ϕ) directly affected by one or more underlying assumptions addressed by model self-evaluation. Mackay and Robinson (1998, 1999) chose terms such as (1) sensitivity, (2) predictability, and (3) synchronicity, which capture the kind of qualities that are of interest in a model sensitivity analysis. Each of these linguistic terms is described using fuzzy memberships, which are defined on a spatial domain and instantiated by parallel simulation of semantically erroneous and semantically corrected representations within the same integrated model. The latter representation is derived by applying ϵ_{ϕ} to the estimate of the semantically erroneous variable. For instance, the patch-level water table position is adjusted to so that fine resolution and coarse resolution spatial models within RHESSysD are interoperable. However, hysteresis, capacitance, and other dynamic qualities within the model may prevent model interoperability given the user's context (α). For instance, the user context, *soil moisture*, results in a map (Figure 3) showing semantic distance between the interpretations provided by the sub-models of RHESSysD. In this case, the user (human or computer) could determine where the models give acceptably close interpretations, in much the same way that Sheth and Kashyap (1992) use a bounded correctness criterion to determine if two objects refer to the same thing.

5 Discussion

Much has been said about the importance of metadata for geographic information systems, and its application to heterogeneous information and data interchange (*e.g.*, Drew and Ying, 1998). For instance, efforts at developing National Spatial Data Infrastructure or international counterparts (*e.g.*, Coleman and Nebert, 1998) and related issues associated with digital libraries (*e.g.*, Tennant, 1998) both emphasize metadata standards. However, these applications do not address semantic integration issues associated with differences in spatial and temporal scales, which must be resolved if simulation models are to become part of the global information infrastructure. Here it has been argued that semantic integration within current integrated models can provide important insight on how to resolve spatial and temporal scale issues as models are combined and moved from application to application.

Semantic integration requires tools for relating objects, simulation model results, and other information sources. One such tool developed here is a measure of semantic proximity. Sheth and Kashyap (1992) also characterize the degree of semantic similarity, or bounded closeness, between a pair of objects using the concept of semantic proximity under fuzzy logic. An alternate approach is taken by the SCOPES system (Ouksel, 1999) in which a Dempter-Schafer probabilistic approach is used to address uncertainty in semantic reconciliation. These approaches are complimentary in that they both recognize the need for context and handling uncertainty in forming semantic linkages between objects. A similar argument is made here in the context of simulation models. We take the view that semantic linkages between sub-models of an integrated model are context-sensitive and must rely both on common knowledge context, such as physical laws, as well as end-user context. However, here the objects in question are generated by the integrated simulation.

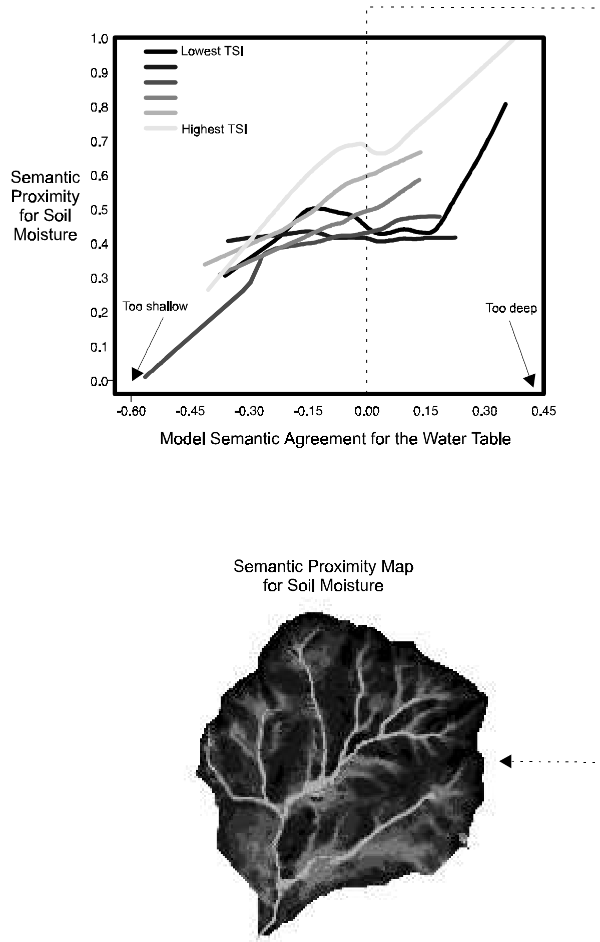

Figure 3. Example of semantic proximity for soil moisture as predicted by the integrated model RHESSysD. The image at the bottom represents a snapshot taken where model agreement is zero, which corresponds to an interoperable interpretation of the water table.

The goal of this approach is to preserve as much end-user and model developer autonomy as possible, in order to make simulation models a viable component of global information systems. A key component of this is providing linguistic terms that are consistent with the end-user's own ontology, while giving the model developer full liberty in making assumptions in their view of environmental systems. Lee and Siegel (1996) argue that presenting answers in a manner consistent with end-user preferences is key to reducing the cognitive effort required of the decision maker to interact with multiple, unfamiliar and dynamically changing sources. Kashyap and Sheth (1997) argue that, where information sources are designed and developed independently, as long as information is consistent within the context of the query of the user, inconsistency in information from different databases is allowable. A similar argument can be made here in the context of environmental models. The example given in the previous section showed that semantic inconsistencies are only critical where they result in an inconsistency within the user's context. For some areas the simulated soil moisture may be quite acceptable given the context.

6 Conclusion and Future Work

It is suggested here that model self-evaluation can preserve model autonomy while resolving certain interoperability problems within integrated environmental models. However, where certain variable dependencies in space and time dictate it may be necessary to present semantic weakness to the end-user. It is suggested here that linguistic terms can present semantic proximity in a way that is meaningful to the end-user. This has the benefit of reducing the cognitive demands on a user who is trying assemble environmental systems knowledge from existing disparate sources, whether they are in the form of current integrated models or in the form of information resources distributed across a global network.

The approach presented here is significant in that (1) semantic reconciliation is query directed, (2) semantic differences are considered in both application and end-user contexts, (3) a multiple criteria reasoning is used in light of incomplete information, and (4) semantic proximity is directly related to a spatial context by presenting it in the form of a map of fuzzy memberships. This semantic integration analysis of environmental models goes beyond existing work in the area of model management and is a step in the right direction for viewing simulation models as information providers to global information systems. However, much remains to be done before the results presented here can be generalized to other environmental models and applications. Some specific areas of further research are:

- Additional empirical work with existing integrated models to determine how well semantic integration can be addressed with a range of user contexts;

- Implementation of environmental models as component-ware that can be integrated as needed to serve specific user contexts; and

- Implementation of "virtual environmental models" that operate in a distributed fashion across a network, preserving individual model and end-user autonomy.

7 Acknowledgments

The work presented in this paper was supported by the University of Wisconsin - Madison Graduate School (Wisconsin Alumni Research Foundation).

8 References

Bennett, D.A. 1997. A framework for the integration of geographical information systems and modelbase management. *International Journal of Geographical Information Science*, 11(4), 337-357.

Bunge, M. 1979. *Ontology II: A World of Systems*, D. Reidel Publishing Company, Boston.

Coleman, J. and D.D. Nebert. 1998. Building a North American spatial data infrastructure. *Cartography and Geographic Information Systems*, 25(3), 151-160.

Drew, P. and J. Ying. 1998. Metadata management for geographic information discovery and exchange. In Sheth A. and W. Klas (Eds.). *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw Hill, 89-121.

Dunn, S.M., R. Mackay, R. Adams, and D.R. Oglethorpe. (1996). The hydrological component of the NELUP decision-support system: an appraisal. *Journal of Hydrology,* 177, 213-235.

Jamieson, D.H. and K. Fedra. (1996). The 'WaterWare' decision-support system for river-basin planning. 1. conceptual design. *Journal of Hydrology,* 177, 163-175.

Kashyap, V. and A. Sheth. 1996. Schematic and semantic similarities between database objects: A context-based approach. VLDB Journal, 5(4), 276-304.

Kashyap V, Sheth A 1998 Semantic heterogeneity in global information systems: the role of metadata, context and ontologies. In Papazoglou M, Schlageter G (eds) *Cooperative Information Systems: Current Trends and Directions*. Academic Press, 139-178.

Keller, R.M., M. Rimin, and A. Das. (1994). A knowledge-based prototyping environment for construction of scientific modeling software. *Automated Software Engineering,* 1, 79-128.

Lee, J.L. and M.D. Siegel. 1996. An ontological and semantical approach to source-receiver interoperability. *Decision Support Systems*, 18, 145-158.

Mackay, D.S., K. Gardels, X. Lopex, H. Foster, and J. Radke. 1996. Interoperability of Geographic Information. University Consortium on Geographical Information Science Research Priority, http://www.ncgis.ucsb.edu/other/ucgis/research_priorities/paper5.html

Mackay, D.S. and V.B. Robinson. 1999. A multiple criteria decision support system for testing integrated environmental models. International Journal of Fuzzy Sets and Systems. (In press).

Mackay, D.S. and V.B. Robinson. 1998. Model self-evaluation and detection of "semantic error" in a spatially explicit ecosystem process model. Proceedings of the Seventh International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Editions, E.D.K., Paris, 588-595.

Mackay, D.S., V.B. Robinson, and L.E. Band. 1994. A knowledge-based approach to the management of geographic information systems for simulation of forested ecosystems. In Michener, W.K., J.W. Brunt, and S.G. Stafford (Eds.). *Environmental Information Management and Analysis: Ecosystems to Global Scales*, Taylor & Francis, London, 515-538.

Oreskes, N., K. Shrader-Freshette, and K. Belitz. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263, 641-646.

Ouksel, A.M. 1999. A framework for a scalable agent architecture of cooperating heterogeneous knowledge sources. In M. Klusch (Ed.). *Cooperative, Rational and Adaptive Information Gathering in the Internet*, Springer-Verlag, In press.

Sheth, A. 1998. Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In Goodchild, M.F., M.J. Egenhofer, R. Fegeas, and C.A. Kottman (Eds.). *Interoperating Geographic Information Systems* (In Press).

Sheth, A. and V. Kashyap. 1993. So far (schematically) and yet so near (semantically). *Proceedings of the IFIP TC2/WG2.6 Conference on Semantics.*

Tennant, R. 1998. Interoperability: The Holy Grail. *Library Journal*, 123(12), 38-39.

van Deursen, W.P.A. 1995. *Geographical Information Systems and Dynamic Models*, Faculteit Ruimtelijke Wetenschappen Universiteit Ultrecht, Ultrecht, Netherlands.

Ventrone, V. and Heiler, S. 1991. Semantic heterogeneity as a result of domain evolution. *SIGMOD Record*, 20, 16-20.

Wesseling, C.G., D. Karssenberg, P.A. Burrough, and W.P.A. Van Deursen. 1996. Integrating dynamic environmental models in GIS: the development of a dynamic modelling language. *Transactions in GIS,* 1(1), 40-48.

Worboys, M.F. and S.M. Deen. 1991. Semantic heterogeneity in distributed geographic databases. *ACM SIGMOD Record,* 20(4), 30-34.

Zeigler, B.P. 1990. *Object-Oriented Simulation with Hierarchical, Modular Models: Intelligent Agents and Endomorphic Systems*. Academic Press, San Diego, CA.