Towards an operational semantics of biological diversity: integrating structure and function in a web-accessible knowledge base

Ferdinando Villa and **Marta Ceroni**

Gund Institute for Ecological Economics, University of Vermont
Until 7/2002: Institute for Ecological Economics, University of Maryland
villa@cbl.umces.edu

Understanding the dynamics of biological diversity requires new approaches and integration between different disciplines. The development of a unitary body of knowledge about biodiversity is made difficult by the fragmentation and lack of interoperability between the diverse - and often informal and simplistic - semantics adopted for data and models. Different representational paradigms require conceptual, as well as technical, integration to enable reorganizing the body of knowledge and enable new ideas, theories, and higher-level paradigms to surface and be tested.

We are developing a web-accessible database where temporally- and spatially- explicit biodiversity data coexist with dynamic modular models, potentially integrating multiple modeling paradigms and automatically enforcing their compatibility. Users of this database will be able to (1) retrieve and explore data and run models from their web browser; (2) compose new models by modifying data, parameters and equations in existing ones using interactive web pages and successive queries of the database; and (3) publish modifications that highlight interesting characteristics of the system (along with corresponding metadata) in the database. We envision a self-evolving, *bulletin board*-like approach to data sharing and collaborative modeling, allowing researchers and managers to share any new insights, and fostering collaboration, synthesis and discovery through the World Wide Web. By means of this approach, this central knowledge repository will bring the power of the open-source development model to the biodiversity research community.

A key point in our project is developing a strong semantic layer that allows common ecological concepts to be mapped directly onto the database contents. As an example, we want to be able to define a species-area relationship as a view on a database, where the spatial, ecological and temporal domains of a set of species are compatible, and abundance data are available for all species. Given a formal definition of this and other common ecological concepts, the database retrieval engine would be able to retrieve all species-area relationships present in the database (possibly with other, optional spatial or temporal constraints), without a need from the maintainers to identify certain sets of observations as such. An object-oriented engine will then allow to operate on the retrieved relationships as defined by the implementation - e.g. fitting common or user-defined species-area mathematical models. We believe that a strong, explicit semantic layer will enable the use of a database as a discovery and synthesis tool, not just as a repository of information.

The XML-based infrastructure used by the underlying software engine, based on the Integrating Modelling Architecture (Villa, 2001; http://www.integratedmodelling.org) enables us to store data and models (modules) adopting an extensible array of representations, and to link them together transparently and automatically. The integration is made possible by explicit cross-cutting abstractions called domains, which represent various aspects of the world (such as time and space) as seen by each module, and are processed by the simulation engine to resolve conflicts related to different scales or representational paradigms. One major innovation brought by the proposed project is that the process of connecting data and model components into higher-level models, operated by the software under guidance of the user, creates a structural specification of the system and the interactions between its components in each domain. This specification can be itself extracted as XML, and represents a well-behaved structure that can be analyzed, optimized, and communicated. Through appropriate post-processing software, we will be able to visualize the system structure as it is modified by the user or the course of the simulation. The resulting representation is an optimal base for advanced applications - e.g., automatic detection of emergent structure - that can be implemented in the system through plug-in extensions.

Users of this database, whose infrastructure will be delivered as open source software and will interact with, but not require, commonly used packages and file formats, will be able to retrieve and link biodiversity data and dynamic models in arbitrary configurations, and visualize or run the resulting system models on a remote server. We will develop a general architecture that can be extended to use, e.g. Individual-based models along with spatially-explicit, process-based models, and provide a significant body of biodiversity data and models as a proof of concept.

References

Villa F., 2001. Integrating Modelling Architecture: a declarative framework for multi-paradigm, multi-scale ecological modeling. *Ecological Modelling* 137: 23-42

Villa F. and Costanza R.. 2000. Design of multi-paradigm integrating modeling tools for ecological research. Environmental Modeling and Software 15:169-177

Villa F. 1992. New computer architectures as tools for ecological thought. Trends in Ecology and Evolution 7:179-183