Challenges in Modelling of Environmental Semantics

Ioannis N. Athanasiadis

Democritus University of Thrace, Xanthi, Greece
ioannis@athanasiadis.info

Abstract. Modelling environmental semantics is a prerequisite for model and data interoperabilty and reuse, both essential for integrated modelling. This paper previews a landscape where integrated modelling activities are performed in a virtual environmental information space, and identifies challenges imposed by the nature of integrated modelling tasks and new technology drivers such as sensor networks, big data and high-performance computing. A set of requirements towards a universal framework for sharing environmental data and models is presented. The approach is demonstrated in the case study of a semantic modelling system for wildlife monitoring, management and conservation.

Keywords: Environmental semantics, Intergrated modelling, Environmental Information Space, Service orientation, Internet of the Things.

1 Introduction

Environmental modeling, almost since its infancy, was challenged with issues of integration and reuse. Today it has become natural to conduct integrated studies by putting together data and models originating from diverse sources. The process starts with the selection of *suitable* models, i.e. capable of producing the desired outputs directly, or outputs that can be easily transformed to the desired ones. Then, model input requirements needs to be matched with data, so that the models can be executed. While this simplification makes it sound as an easy task, in the contrary the reality is very challenging. This process is never a two-step action, rather an on-going, iterative process: data limitations have an impact on the models chosen, and model performance drives the needs for additional data sources. At the same time, questions to be answered change with the better understanding of the system, so that more aspects are covered: the better we understand the system behavior via simulations the more we change it. In this respect, scientists performing integrated modeling are challenged to develop skills that span from tedious data reformatting to advancing science, by creating new models. Integrated modeling is challenged with developing methodologies that manage with the inherit properties of environmental data and models.

Environmental data are spatiotemporally referenced, but (more importantly) uncertain to some degree, as they inherit the measurement instruments' failures, biases and noise [1]. At the same time, environmental data is a resource in scarcity. Already in Agenda 21, it was highlighted that "the gap in the availability, quality, coherence, standardization and accessibility of data between the developed and the developing

R. Denzer et al. (Eds.): ISESS 2015, IFIP AICT 448, pp. 19–25, 2015.

world has been increasing, seriously impairing the capacities of countries to make informed decisions concerning environment and development." [2]

Today, we experience the lack of information not only in the developing countries, where limited data records are available, but also in the developed ones, as we are flooded with data, which are not universally accessible. Environmental data are often hidden in silos, encoded with poor standards, in legacy systems, and some times are not available digitally, or human intervention is needed to access them. Issues of copyright and licensing, though changing fast, still limit open access to environmental data. Despite the abundance of data available still we need scientists to scout for data that are needed for integrated studies.

Environmental models inherit the complexity, uncertainty, scaling, and integration qualities from the physical world [3], which are observed as characteristic properties of the environmental systems. Rizzoli and Young in [4] summarized environmental systems as heterogeneous, spatiotemporal dynamic systems, with stochastic and periodic components. Denzer (2005) [5] to overcome the problems in environmental model integration insisted on model abstraction, communication and generality as three essential tests for model integration. Undoubtedly, most models today wouldn't pass those tests.

Model implementations today are poorly designed and documented, as they have been originally developed for single, or limited use. Model reuse, composition and chaining via workflows are characteristics that we have never designed for. Furthermore, one needs to consider that when an environmental model is encoded in a programming language, new limitations are introduced compared to the original modeling assumptions. Hardly ever can these assumptions be represented directly in the implementation language of choice; on the contrary, this knowledge resides with the modelers [3].

Both data and models encode domain knowledge that resides with the specialists. However this knowledge often is not accessible, and integrated modeling teams need to establish contact with original data and model producers to be trained to use them properly. Undoubtedly, we have not reached a level where data and models come with such a detailed documentation so that third-party scientists can reuse them soundly, or detailed meta-information so that machines can invoke them directly. We are still far away from the vision of a common environmental information space (Figure 1), where agencies, organizations and the public will have unhampered, universal access to environmental data and models.

2 Semantics for an Environmental Information Space

Common information spaces have been realized in other application areas (ie retail, banking, entertainment, travel, etc), so one could argue that it is a matter of time or resources to happen for environmental information. However, this is not the case due to the **subjective** nature of environmental information. In contrast to other areas, both data and models in the environmental sector are *subject to interpretation*. In the case of data collection, attributes measured, instruments used, sampling methods and quality

check procedures depend on the particular goal of the specific study. Have the goals been different; one may have selected different equipment or applied different methodologies, which will have led to other results. The same holds for modeling, as theory, scale and boundaries depend on the problem definition. For such reasons, model integration and reuse in a common environmental information space needs to allow for *interpretation*. There is no universally agreed view of environmental information, which means that we need interpretations relevant for an individual, a project or a community.

Semantic modeling has been proposed as a remedy for overcoming longstanding issues of model integration. In our previous work with Villa and Rizzoli [7] we identified two approaches to semantic modeling. In the *mediation* approach, formal knowledge is the key to automatic integration of datasets, models and analytical pipelines. The next step, applied experimentally at this stage, is the *knowledge-driven* approach, where the knowledge is the key not only to integration, but also for overcoming scale and paradigm differences, and automated knowledge discovery.

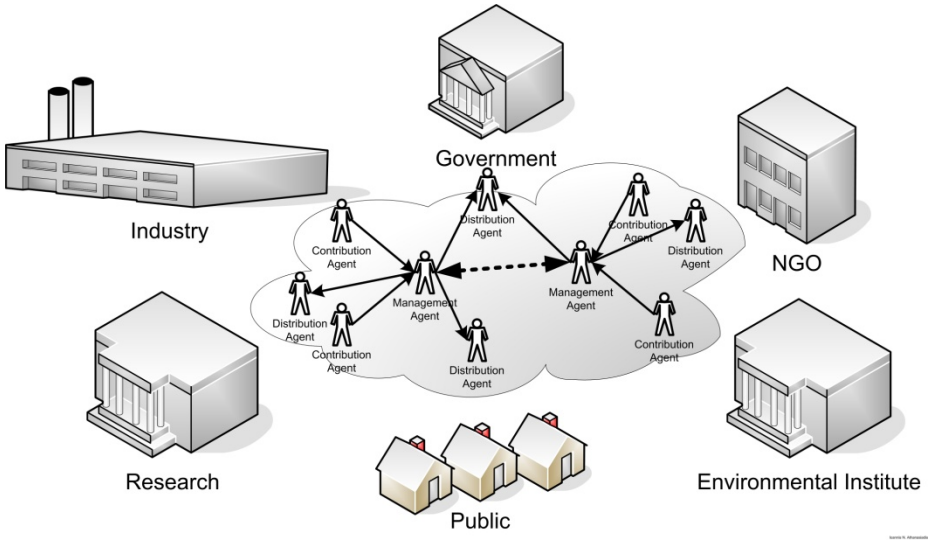

Fig. 1. A vision of a common environmental information space (Figure from [6])

Today, more than ever we are in need for developing common environmental information spaces that enable integrated modeling, following a sharing resource model (Figure 1). Each peer offers data or models, and others are able to discover and reuse them. The prime requirement of such an information space is the need for subjective interpretations: The same data or models can be interpreted differently for different studies. In the mediating approach, the challenge is for semantic annotations that allow for subjectivity. While there have been significant efforts to build domain ontologies by several projects, there was limited take up by broader communities. Apart from a few very basic nomenclatures, the rest of the domain ontologies I have used (or developed) were in one or another way biased by the problem at hand.

The second key requirement for such a common environmental information space is the *transferability* of scientific workflows. We have experienced times and again how difficult it is to perform the same study even in a nearby location: Data sources are missing, models do not converge, and corrective actions or new assumptions are needed. The major problem here is that the expert knowledge is hidden in model implementation or data archives, and our tools are not capable for manipulating our sources. Expert intervention is needed to "*adjust data*" and "*turn model knobs*". A semantically-aware common environmental information space needs to make such dependencies explicit and offer tools to match data offerings with model requirements.

Additionally, a common environmental information space for integrated modeling needs to:

a. Overcome obstacles of syntactic interoperability, by offering plug and play services for transforming data sources
b. Allow for data and model substitution, to enable model comparison in scientific workflows
c. Offer uniform services for output visualization, to allow for less engagement in producing visualizations
d. Document results provenance, ensuring the transparency of results
e. Allow for uncertainty quantification and error propagation
f. Allow for sensitivity analysis

Today, a common environmental information space is further challenged by the Internet of the Things: In the years to come we expect an abundance of sensory data to become available at very low cost, at real near time, over the Internet. This has already started to transform our view on performing local studies, engaging with communities and employing participatory methods for data collection. This will change integrated modeling methodologies, as more data will be around, but at the same time it will raise the bar for discovering such information, annotating them and evaluating their added value. A common environmental information space needs to hook up to sensor networks and allow models not only to run again as new data arrive, but to adapt as conditions change.

Another important factor that challenges our view on integrated modeling is the raise of high performance computing and the technologies for manipulating big data. Hardware acceleration and virtual computing infrastructures already allow massive simulations at a very large scale. However, still there is an entry barrier for making such computing infrastructure available. A common environmental information space needs to provide with seamless access to virtual computing infrastructures.

3 Case Study

In the following, I present a case study where we try to meet some of the challenges of integrated modelling with sensor data using semantic technologies. Based on a Greek NGO experience in large carnivores conservation in the mountain ecosystems of northern Greece, we built a generic architecture for wildlife information fusion,

sharing and reuse. The ALPINE wildlife modeling system (hereafter, ALPINE for short) is a semantic modelling system for wildlife monitoring, management and conservation. ALPINE aims to demonstrate how live streaming data from animal tracking sensors can be effectively combined with geo-statistical analysis models, in order to assess habitat suitability, and to quantify the risks of wildlife interaction with man-made infrastructures. The overall system architecture (Figure 2) involves three layers of services, and is currently under development using Thinklab [7], the semantic modelling infrastructure of ARIES [8].

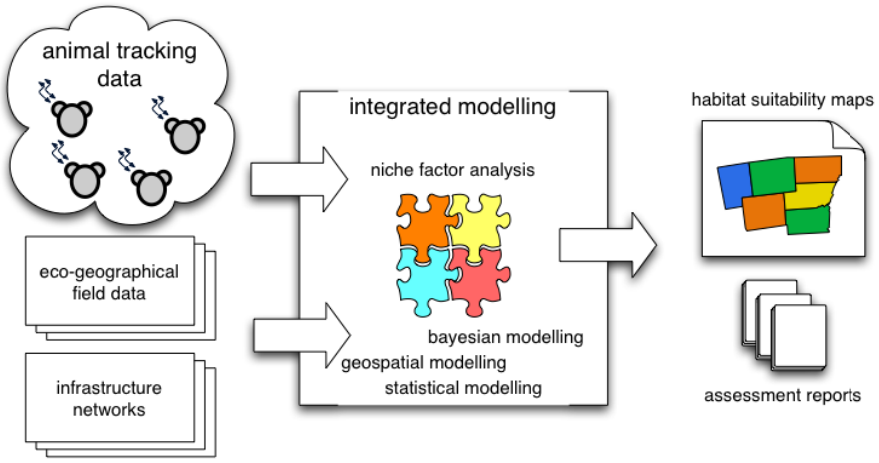

Fig. 2. The ALPINE architecture for wildlife monitoring (Figure from [9])

The first service layer deals with making available environmental data, originating either directly from sensors, or public and private archives. Animal tracking data from collars and eco-geographical field data and infrastructure networks are made available to a common environmental space. The semantic modelling system smooths out technical details for retrieving and transforming data, and also allows for interpretations tailored to the specific modelling exercise. The ALPINE system simplifies access via using open data protocols and making data discoverable through rich annotations. Open Geospatial Consortium standards have been employed to offer syntactic interoperability: Sensor Observation Service [10] for sharing sensor data, and Web Coverage Service [11] for datasets of geographical nature. Both field data, collected by ALPINE sensors (i.e. from GPS/GSM collars), and background information are annotated with problem-specific semantics, which offer interpretations for the particular problem and allows data to be matched to models.

Second, the integrated modeling layer employs statistical, geospatial and Bayesian models for ecological niche factor analysis. Models are annotated with problem-specific semantics and made available to a common environmental space. Thinklab semantic modeling engine allows for chaining models in scientific workflows, substituting models with alternatives, and feeding models with data to produce results. Specifically, three kinds of

models are made available through the ALPINE system: Geospatial models allow operational interpretations of spatial sources and are typically used for creating derived information from original data, as buffering functions and density analysis. Bayesian models are employed for building probabilistic models in order to incorporate causal associations from evidence. For a more detailed discussion on Bayesian modeling for ecological risk assessments see [12]. Last, Ecological Niche Factor Analysis (ENFA) is a statistical procedure that uses only presence data, suitable to compare distributions among spaces that a population has a reasonable probability to occur using eco- geographical variables and the global space [13]. The ALPINE integrates seamlessly these three kinds of models in a platform in order to enable scientists to perform their assessments.

Third, the presentation layer generates maps and reports with the system results. Typically scientists spend adequate amount of time in order to analyse their results and post-process them. The ALPINE system will incorporate such aspects in the workflow, so that maps and reports are generated, as new data arrive in the system and assessments are updated. For this we employ reusable templates that will incorporate model results.

The ALPINE system is intended for scientists who aim to answer questions related to habitat suitability and wildlife-human interactions. It enables scientists to hook up sensor data streams coming live from sensors with geographical information and build scientific workflows to support integrated modeling studies. The ALPINE system tackles some of the semantic challenges for incorporating sensors in integrated modelling studies: The Thinklab modelling engine of ALPINE (a) minimizes human involvement in data preprocessing and manipulation, especially as new data arrive from sensors; (b) makes easier to re-run models, as new data arrive from sensors; and (c) provides tools for exporting results in different formats.

4 Epilogue

This paper aimed to preview some challenges for integrated modeling through a common information space of semantically shared environmental data and models. I believe we are close in realizing such a vision. Many of the building blocks are already in place. We have several success stories for standardizing nomenclatures, offering data as services through long-term archives, making model available as services and enabling model composition and execution in local or remote infrastructures. At the same time we are trapped with legacy software and institutional problems that do not allow such a vision to come true. Another significant part I didn't touch in this paper is the human side of the problem. In the current academic and scientific system there are very little incentives for building a sharing culture, which is a prerequisite for a common information space for integrated modeling.

Acknowledgements. Research in this study has received funding by European and national funds from NSRF 2007-2013, OP Competitiveness and Entrepreneurship, Cooperation 2011, in the context of the ALPINE project (grant 11SYN-6-411). I am grateful to Prof Ferdinando Villa for the insightful discussions that led me to the views presented here.

References

1. Rizzoli, A.E., Athanasiadis, I.N., Villa, F.: Delivering environmental knowledge: a semantic approach. In: Proc. 21st International Conference on Informatics for Environmental Protection (EnviroInfo 2007), pp. 43–50. Shaker Verlag, Warsaw (2007)
2. UN Earth Summit: Agenda 21. Department of public information, United Nations, Rio de Janeiro, Brazil (1992)
3. Athanasiadis, I.N., Villa, F.: A roadmap to domain specific programming languages for environmental modeling: key requirements and concepts. In: Proc. 2013 ACM workshop on Domain-Specific Modeling, pp. 27–32. ACM (2013)
4. Rizzoli, A., Young, W.: Delivering environmental decision support systems: Software tools and techniques. Environmental Modelling & Software 12, 237–249 (1997)
5. Denzer, R.: Generic integration of environmental decision support systems - state-of-the-art. Environmental Modelling & Software 20, 1217–1223 (2005)
6. Athanasiadis, I.N.: Towards a virtual enterprise architecture for the environmental sector. In: Protogeros, N. (ed.) Agent and web service technologies in virtual enterprises, pp. 256–266. Information Science Reference, Hershey (2007)
7. Villa, F., Athanasiadis, I.N., Rizzoli, A.E.: Modelling with knowledge: a review of emerging semantic approaches to environmental modelling. Environmental Modelling and Software 24, 577–587 (2009)
8. Villa, F., et al.: Thinklab software repository (2013)
9. Villa, F., Bagstad, K.J., Voigt, B., Johnson, G.W., Portela, R., Honzak, M., Batker, D.: A methodology for adaptable and robust ecosystem services assessment. PLoS ONE 9, e91001 (2014)
10. Athanasiadis, I.N., Villa, F., Examiliotou, G., Iliopoulos, Y., Mertzanis, Y.: Towards a semantic framework for wildlife modeling. In: Marx Gomez, J., et al. (eds.) Proc. 28th International Conference on Informatics for Environmental Protection (Enviroinfo 2014), pp. 287–292. BIS-Verlag, Oldenburg (2014)
11. OGC: Sensor Observation Service, Open Geospatial Consortium Standard (2007)
12. OGC: Web Coverage Service, Open Geospatial Consortium Standard (2012)
13. Pollino, C.A., Woodberry, O., Nicholson, A., Korb, K., Hart, B.T.: Parameterisation and evaluation of a bayesian network for use in an ecological risk assessment. Environmental Modelling & Software 22, 1140–1152 (2007)
14. Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N.: Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83, 2027–2036 (2002)