In **Proc. 21st International Conference on Informatics for Environmental Protection: EnviroInfo 2007**, (Hryniewicz, Olgierd, Studziński, Jan, and Romaniuk, Maciej, Eds.), Shaker Verlag, September 2007, pp.43-50.

Delivering environmental knowledge: a semantic approach

Andrea E. Rizzoli*, Ioannis N. Athanasiadis*, Ferdinando Villa**

Abstract

Environmental informatics delivers techniques and tools for archiving and processing environmental data. The advent of the Internet had positively affected the availability and ease of access to large and diverse environmental databases, distributed all over the world. On the other hand, similar progress has not been matched by the availability of models and algorithms able to process these data, mostly because of the lack of standards in the annotation of the characteristics of environmental models. In this paper we advocate the need for the semantic annotation of environmental "knowledge", encompassing models and data. The slow, but steady, introduction of the Semantic Web and the widespread use of ontologies for semantic annotationRizzoliAV_ will allow environmental informatics to cover the gap in the access and usability of models and algorithms for environmental data processing.

1. Introduction

Environmental informatics (EI) has been defined by Avouris and Page (1995) as the 'study and development of adequate techniques for collection, storage, retrieval and processing of complex environmental data'. A definition that stands the test of time, since it is still valid after more than 10 years. Yet, this definition does not tell us how environmental informatics is performing with respect to its declared aims.

While the above definition is crisp clear, the only ambiguous part resides in the adjective 'adequate'. The adequacy of EI techniques is subjective and relative to the kind of application and to the intended end-user. The question is: are our techniques adequate?

The answer is, as anticipated, subjective, and probably it depends on the perspective we look at the problem. From the point of view of delivering support in decision making within the specific context of given situation, the answer is mostly positive; for instance, Turon *et al.* (2007) report on the successful implementation of a DSS for the management of constructed wetlands for water treatment, while a broader list of recent efforts in the design and implementation of environmental DSS can be found in Matthies *et al*. (2007). On the other hand, if we investigate the reusability of EI techniques, we are disappointed in discovering that we can partly reuse our collections and storages of environmental data, but when it comes to reusing techniques for "processing complex environmental data", the success rate is extremely low.

Collecting, storing and retrieving environmental data is performed thanks to database techniques, while *processing* environmental data pertains to the field of modelling: data are used to generate information. A model can be as simple as a database query, but it can also be a complex mathematical algorithm, solving a set of partial differential equations over a spatial and temporal domain. Thus, we see why we fail in reusing techniques for processing environmental data: databases have been intensively used and standardised over the years; model bases, despite having been introduced nearly 20 years ago (Guariso and Werthner, 1989), have not really taken off. Being unable to access already available models also negatively affects the ability to reuse them and to combine models across disciplines and domains, as required by sustainability studies and integrated assessments (Denzer *et al.* 2005).

* IDSIA, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Galleria 2, 6928 Manno, Switzerland. Email: {andrea, ioannis}@idsia.ch.

** Gund Institute for Ecological Economics, University of Vermont, VT, USA. Email: ferdinando.villa@uvm.edu.

In the remainder of this paper we examine the reasons behind the failure of model bases and we propose a solution, which relies on a semantically mediated access to distributed information.

2. A brief review of EI tools

The techniques developed in the environmental informatics field are then implemented and find their incarnation in an array of software tools, platforms and environments. We can distinguish among:

- data storage infrastructure software,
- data processing infrastructure software,
- environmental software development platforms and frameworks,
- end-user applications.

The main tools available in the storage infrastructure software category are *databases*. Basically environmental databases differ from non-environmental ones only for their content, and there are no major structural differences. However, there are some conceptual differences. Environmental databases typically contain scientific measurements, as the result of observing natural phenomena. As such, environmental data are spatiotemporally referenced, but (more importantly) uncertain to some degree, as they inherit the measurement instruments' failures, biases and noise. These two points, along with the documentation of the observation process are the critical characteristics of environmental data that environmental informatics need to deal with, and differentiate it with contemporary business-related data management. To give an example, an environmental database with climate data does not simply contain time series of sensor recordings, but it also needs to capture spatio-temporal references, units and dimensions of the measurements, the type and the accuracy of the sensor device, and a specification on how the measurements have been taken (i.e. at the ground level, under shadow, etc)

Also in the category of data storage we find *model catalogues* and *model bases* (see for instance the GAIA model catalogue[1], and the EPA model catalogue[2]). Note that sometimes the term "model" defines the whole computer application providing an implementation of a given (mathematical) model. In the following, the term model will always refer to the specific formal representation and not to the software application. Provided this distinction, a model catalogue is simply a listing of the characteristics of a computer model, mostly in textual format, possibly providing access to the model executable or source code; on the other hand a model base delivers additional search capabilities, and access to the mathematical structure of the model (Benz *et al*. 2001). Note that opening the access to the model structure also requires being able to formally represent it.

The data processing infrastructure software includes *GIS*, *expert systems* and *case-based reasoning systems*, software for *statistical analyses*, data *classification* algorithms, *simulation* tools and *optimisation* algorithms. It is a very wide software class and, again, there is not really a unique environmental flavour to it, but it is more the kind of application that distinguishes such software as "environmental". In the case of expert systems and case-based reasoning systems, the rules and constraints, which are used to make inferences on environmental data, are often called the knowledge base.

Environmental software development platforms and frameworks are meta-tools, analogous to integrated development environments for developing standard software applications, such as Eclipse or Visual Studio .NET. Rizzoli et al. (in press) discuss the case of environmental integrated modelling frameworks, specific platforms which target modelling as their main output, and advocate for a number of specific requirements that distinguish them from standard software development platforms. Among environmental software development platforms we can list TIME (Argent, 2004), JAMS (Kralisch and Krause, 2006). Some frameworks focus on specific aspects as model linking: an example is OpenMI (Gregersen *et al.* in

[1] http://www.ess.co.at/GAIA/models.html

[2] http://www.epa.gov/epahome.models.htm

press) that also proposes a standard interface for model linking. Environmental software development platforms and frameworks are used to deliver end-user applications, providing advanced software engineering techniques to facilitate the software development process in all its stages.

End-user applications basically include all kind of environmental software applications targeted at the different end-users, from policy makers up to environmental interest groups. A very relevant class of end-user applications are *environmental decision support systems* (Guariso and Werthner, 1989). It is to be remarked that only a very minor part of end-user applications are developed by means of specific environmental software development platforms.

3. Engineering the software development process in EI

A main concern of the EI community is (or should be) the inefficiency of the software development process. Good software engineering practices are often disregarded and there is a need to address this issue. In most cases, end-user applications are developed *ad hoc* with a very low degree of re-use of existing tools, models, and databases.

All hints point to problems in the design of end-user applications, where data, models and algorithms are often entangled and good software design principles are disregarded. Re-using an existing application in a different context requires major code re-factoring and very often it is less costly to re-implement from scratch. A true separation of concerns is thus required. Models, data and algorithms should be kept separate in order to foster reusability, maintainability, transparency and access to documentation (Rizzoli et al 1998). For instance, the model equations of a dynamic model can be kept separate from the numerical algorithm solving the systems of differential equations. In a similar way, the rules in a knowledge base can be stored independently of the inference algorithms operating on them. The clear advantage of this approach is the ability to re-use the model (the equations, the rules), when the tool (the numerical integrator, the inference engine) changes.

Such a separation of concerns can be achieved thanks to component-based software engineering (CBSE), a modern approach to software engineering (Szyperski *et al.* 2002). In the context of environmental informatics, a software development platform could see manipulate storage and data processing components to assemble and deliver end-user applications (Rizzoli and Argent 2006). Among data processing components we can list computer models, which publish their interface (Donatelli *et al.* 2006).

Many commercial modelling and simulation tools display a component-oriented approach for model building (e.g. Simulink[3], Labview[4], Extend[5], Arena[6]). The modeller can access a wide palette of basic model components, which can be linked together in order to build more complex models that in turn can be stored back in the model base.

4. The limits of component-based software for EI

The component-based approach to modelling has proven very successful in different domains, from electronics (e.g. SPICE[7]) to mechanical engineering (e.g. Modelica[8]), chemical engineering (CAPE-OPEN[9]) and many other engineering sectors. Yet, its adoption and use in the environmental science sector

[3] http://www.mathworks.com/products/simulink/

[4] http://www.ni.com/labview/

[5] http://www.imaginethatinc.com/

[6] http://www.arenasimulation.com/

[7] http://bwrc.eecs.berkeley.edu/Classes/IcBook/SPICE/

[8] http://www.modelica.org/

[9] http://www.colan.org/

is still lagging behind. While we could point at first to great diversity of modelling approaches and techniques in the environmental sector, we must also acknowledge that environmental systems are "live" systems of considerable dynamic complexity: a resistor can be excellently approximated by Ohm's law, but an ecosystem cannot be boxed in a closed formula. Thus, packaging and delivering ready-to-use software components for environmental modelling is no easy job.

One may object that as we steer towards environmental engineering, for instance hydrology, it's easier to find standard modelling approaches (e.g. the linear cascade reservoir model for rainfall-runoff), while in environmental modelling the same phenomenon can be described by different model paradigms, from mechanistic models, to physical based models to statistical models.

Also, environmental models are often used in integrated assessments and sustainability studies, which are inherently trans and cross-disciplinary. As a consequence, the modelling domain is extremely wide, including social and economic modelling. The interchangeability of model components is therefore limited when different paradigms are to be used in the same integrated model. Dealing with a number of models interacting according to complex patterns also raises problems related to a consistent and thorough representation and propagation of uncertainty, which affects environmental models at all levels, in data, in the structure and in the parameters. Current research efforts are targeting these problems with an array of methodologies, from modern sensitivity analysis (Saltelli *et al.* 2000) to inverse quantitative analysis (Ravalico *et al*. 2006).

As a consequence, the knowledge on how to build a complex integrated environmental model is very often spread across a number of disciplines and domains. It is quite unique that a single scientist owns all this knowledge and it is even more difficult that it can be stored on a single machine.

The advent of the Internet has drastically changed the way we search for information but also the way we store it. We can now rely on information, tools and software solution, which are not installed on our computers but are on the Net. The problem is that information is owned by a number of different actors and information is encoded in a myriad of alternative ways. Having a single environmental model base is a utopia; what is nowadays realistic is to think of the Internet as a distributed environmental model base.

5. Ontologies and the semantic web for environmental modelling

The Internet and its powerful search engines let us access an impressive number of environment-related resources. The search engines are getting better and better at ranking the most relevant result first, but still we need the human intervention to verify and prepare the retrieved information for its further perusal. For instance, we want to find a computer model for simulating groundwater contaminant transport. Searching "groundwater simulation" we hit 1'190'000 results with Google, and narrowing the search to "groundwater contaminant transport simulation" yields "only" 747'000 results. Among these only a small number, less than ten, are links to computer models. Even if we assume that we can find the ten relevant hits, how do we operate these models? Each one has its inputs and outputs requirements, runs on a specific platform, and requires the calibration of possibly different parameter sets.

The World Wide Web as it is now is designed for human consumption, not for being machine processed. The semantic web initiative proposes to change this situation, making the web easily to be parsed and interpreted by computers, as described in the seminal paper by Berners-Lee et al. (2001). The aim of the Semantic Web is to enable the automatic and meaningful interoperation of business-to-consumer and business-to-business applications. Accessing a remote environmental data or model base can be seen as a typical business-to-consumer application, while two environmental models, sitting on different computers in different locations, running and exchanging data, can be seen as a typical business-to-business application.

So far, environmental models have been treated as stand-alone applications, typically deployed in isolation, in strictly configured machines and stringent data requirements. Data preprocessing and manage-

ment, model orchestration and composition has been thought of as modellers' assignments, that should take place in code, and can be achieved by putting enough programming efforts. However, practice has shown, that linking models is not a technical task that affects only software, but primarily it is a modeling task, that requires deep knowledge of the models to be linked. This practice has led the community to fail in sharing effectively models and data. From an academic perspective, several models have been re-developed from scratch, though similar models already existed, for simulating the same phenomena, as the earliest models were not available for inspection, not designed for accepting extensions, or were not accessible for performing comparative runs. This ended up in several software implementation variants for the same models, all of which are tightly coupled with data. Undoubtedly, software abstraction, interoperability and reuse are properties that have been often evangelized, but have never really been taken into account in the modeling, software design, development, and deployment process. Similar is the situation with the environmental datasets, which usually reside in legacy systems, and accessing them requires low-level data processing, by specialists who are aware of concealed assumptions and particular conditions.
Our vision is that the environmental software community has to take a major step ahead; by recognizing the need for shifting the current paradigm, to a new era, where environmental software and data will eventually become shared assets, available as services. The Internet is the infrastructure for realizing this vision, and semantics is the enabling technology, as we discuss below.
Data format standardizations with rich semantics are the key for enabling advanced computer-based services in the semantic web. To give an example, the whole Geospatial Web experience (Scharl, 2007), and its defenders, as Google Earth[10], Yahoo Maps[11] and NASA World Wind[12], are enabled on top of the Geography Markup Language (GML), a standard developed by the Open Geospatial Consortium, and released by ISO (Cox et al, 2004). Similar are the needs for environmental information in general, and there are several ongoing efforts on defining standards for sharing data related to the natural environment and opening them up in the Internet (also discussed in Athanasiadis 2007). These include the work of the US Environmental Data Standards Council[13], released in January 2006, along with the standards developed gradually since 1994, by the European Environment Information and Observation Network (EIONET[14]), and those of the Ecological Society of America (VEGBANK[15]). Also, Food and Agriculture Organization (FAO) of the United Nations has recently made its thesauri of food and agricultural terms, publicly available through the AGROVOC web services[16].

These efforts, along with the work of several research projects all around the world, as SPiRE[17] (semantic prototypes in environmental informatics), SEEK[18] (science environment for ecological knowledge), SEAMLESS[19] (system for environmental and agricultural modeling: linking European science and society), ARIES[20] (assessment and research infrastructure for ecosystem services) are shaping the way for publishing environmental datasets in standard formats on the Web 2.0. In the years to come, we expect a virtual flood of environmental data, as public authorities will start making their records available online. This information will be annotated with rich semantics, so that intelligent agents will be able to manipulate complex queries like "*What is the evolution of trout population in the lake of Lugano?*", and result not

[10] http://earth.google.com/
[11] http://maps.yahoo.com/
[12] http://worldwind.arc.nasa.gov/
[13] http://www.envdatastandards.net/
[14] http://www.eionet.europa.eu/
[15] http://www.vegbank.org
[16] http://www.fao.org/agrovoc/
[17] http://spire.umbc.edu/
[18] http://ecoinformatics.uvm.edu/projects/seek.html
[19] http://www.seamless-ip.org
[20] http://ecoinformatics.uvm.edu/projects/aries.html

only to spatiotemporal visualizations, but also to classifications according to ecological indexes, as species taxonomies, GMOs, etc. In the next section we discuss how this new breed of environmental models and data should be deployed.

6. Model interoperability and scientific workflows

Different authors (among others: Medeiros *et al*. 2005, Bowers and Ludäscher 2005, Lee *et al*. 2007) have envisioned a near future where the modeller will be able to compose different environmental models and data, assisted by the semantic web, which will check for the compatibility of the model linkages (in terms of units dimensions etc), the sound application of the model to the given spatio-temporal context, and it will also select the right algorithms and tools to process these data and models. In such environments, models are deployed as services, and service orchestration is the key for composing integrated models. Building upon this paradigm, more complex models can be deployed as services. Today, there are a variety of technological solutions available for deploying models in a service-oriented manner. These include web services, software agents and grid computing

The Semantic Web, may serve as the infrastructure for deploying an open environment, where diverse peers formulate virtual enterprises, i.e. constellations that provide environmental data and models as services. The members of the enterprise through a privilege management authority may grant access to data, models and computing power. In such an open infrastructure, as the semantic web, environmental agencies, research institutes, NGOs, the industry and the public are enabled to share models and data, treating them as a common good (Figure 1). *The virtualization of a collaborative environment is essential for treating environmental information as a common asset that is shared among peers, instead (of treating it) as a resource in scarcity that peers strive for* (Athanasiadis 2007).

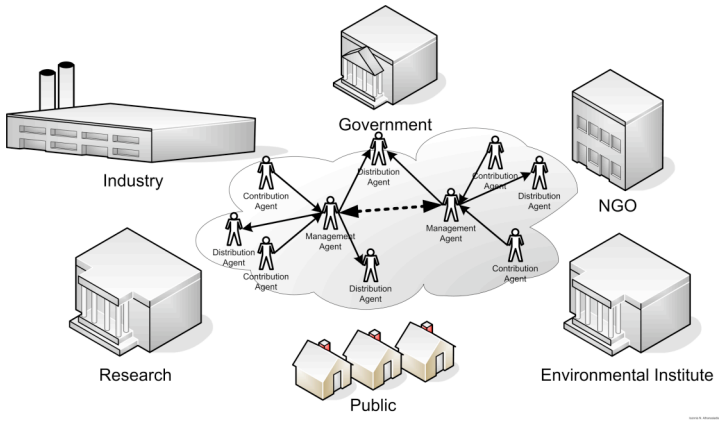

Figure 1: In an open infrastructure, as the semantic web, environmental agencies, research institutes, NGOs, the industry and the public are enabled share models and data, treating environmental information as a virtual asset.

However, in order to achieve interoperability across these services, it is required to agree upon common specifications about the service interface in both technical terms (i.e. methods and tools for publishing services, service registries and lookups, service invocation etc) and term of content (model input and output requirements, parameter specification). On the first front, there are major efforts led by the IEEE-FIPA[21] for specifying agent platform interaction protocols, the W3C[22] consortium for (semantic) web serv-

[21] http://www.fipa.org/

[22] http://www.w3c,org

ices (WSDL, OWL-S), and the Open Grid Forum[23] and the Globus Alliance[24] for grid computing technologies. Service-oriented technologies attract major interest by the software industry; so several tools are available to enable these technologies. What remains the major challenge is the second front, the standardization of the content that these services will exchange. And this is what we envision that the EI community needs tackle the coming years. The specification of shared conceptualizations of environmental data structures, which will turn environmental information into a virtual asset and enable the efficient orchestration of environmental data and model services. An approach similar to the one of ebXML[25] (Electronic Business using eXtensible Markup Language), that enables enterprises of any size and in any geographical location to conduct business over the Internet, is required for environmental data. Building upon existing standards, as GML[26] for geographical information and WSDL[27] for web services, extensions for specific target areas may arise, i.e. for crop-growth modeling, water management, etc. Then the following step will be to embrace all these into a mesh-up infrastructure, forming a computing platform for Environmental Informatics software development and deployment.

7. Conclusions

Environmental informatics, as well as bioinformatics, is at the forefront of research in the application and deployment of new and emerging software and computer technologies. This is happening because of its immense appetite for large databases and the underlying complexity of the physical phenomena to be modelled. More and more environmental data are being made accessible on the internet, yet we are facing two major problems: data are often locked in databases, tightly dependent on their private schemas and database engine; moreover, finding the right environmental models to process these data is more an art than an engineering task. We believe that these two obstacles to further development of environmental informatics can be overcome thanks to the adoption of the semantic web approach and of its standards. We are experiencing the use of ontologies for model linking and to provide semantic annotation of large and unwieldy environmental data sets. This approach paves the way for a near future of environmental data and modelling resources, distributed on the net and seen as a seamless computational grid.

Bibliography

Avouris N.M. and Page B., (1995) *Environmental Informatics, Methodology and Applications of Environmental Information Processing*, Kluwer Academic Publishers, Dodrecht.

Turon, C., Comas, J., Alemany, J., Cortés, U., Poch, M. (2007) Environmental decision support systems: A new approach to support the operation and maintenance of horizontal subsurface flow constructed wetlands, *Ecological Engineering* 30 (4), pp. 362-372.

Matthies, M., Giupponi, C., Ostendorf B. (2007) Environmental decision support systems: Current issues, methods and tools, *Environmental Modelling & Software*, 22 (2), pp. 123-127.

Guariso G. and Werthner H. (1989) *Environmental Decision Support Systems*. John Wiley & Sons.

Denzer R., Riparbelli C. Villa M., Güttler R. (2005) GIMMI: Geographic Information and Mathematical Models Inter-operability. *Environmental Modelling & Software*, 20 (12), pp. 1478-1485.

Benz J., Hoch R., Legovic T. (2001) ECOBAS - Modelling and documentation, *Ecological Modelling* 138 (1-3), pp. 3-15.

[23] http://www.ogf.org/

[24] http://www.globus.org/

[25] http://www.ebxml.org/

[26] http://www.opengis.net/gml/

[27] http://www.w3.org/TR/wsdl

Rizzoli A.E., Donatelli M., Athanasiadis I.N., Villa F., Huber D. (to appear) Semantic links in integrated modelling frameworks, *Mathematics and Computers in Simulation*.

Argent R.M. (2004) An overview of model integration for environmental applications—components, frameworks and semantics, *Environmental Modelling & Software*, 19(3), pp. 219–234.

Kralisch S. and Krause P. (2006) JAMS – a framework for natural resource model development and application. In A. Voinov, A.J. Jakeman, and A.E. Rizzoli, editors, *Proceedings of the iEMSs Third Biennial Meeting, "Summit on Environmental Modelling and Software"*, Manno, Switzerland.

Gregersen J., Gijsbers P., Westen S. J. P. (in press) OpenMI: OpenMI – Open Modelling Interface. *Journal of Hydroinformatics*.

Rizzoli A.E., Davis J.R., Abel D.J. (1998) Model and data integration and re-use in environmental decision support systems. *Decision Support Systems*, 24 (2) pp. 127–144.

Szyperski C., Gruntz D., Murer S. (2002) *Component software - beyond object-oriented programming*, ACM press, New York.

Rizzoli A.E. and Argent R.M. (2006) Software Systems: Platforms and Issues for IWM problems, in: C. Giupponi, A.J. Jakeman, D.Karssenberg, M.Hare (eds.) *Sustainable Management of the Water Resource: an Integrated Approach*, Edward Elgar, pp. 324–346.

Donatelli M., Carlini L., Bellocchi G. (2006) A software component for estimating solar radiation, *Environmental Modelling & Software*, 21 (3), pp. 411-416.

Saltelli A., Tarantola S., Campolongo F., Ratto M. (2004) *Sensitivity analysis in practice - a guide to assessing scientific models*. John Wiley and Sons, Chichester.

Ravalico J.K., Maier H.R., Dandy G.C., Norton J.P., Croke B.F.W. (2005) A comparison of sensitivity analysis techniques for complex models for environment management. In: Zerger, A., Argent, R.M. (Eds.), *MODSIM 2005 International Congress on Modelling and Simulation*, Modelling and Simulation Society of Australia and New Zealand, December 2005. Melbourne, Australia, pp. 2533-2539.

Berners-Lee T., Hendler J., Lassila O. (2001) The Semantic Web, *Scientific American*, 284, pp. 34-43.

Shadbolt N., Hall W., Berners-Lee, T. (2006) The Semantic Web revisited. *IEEE Intelligent Systems*, 21 (3), pp. 96-101.

Scharl, A. (2007). Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories. In Scharl, A. & Tochtermann, K. (ed.) The Geospatial Web: How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society. Springer-Verlag, , 3-14

Cox, S., Daisey, P., Lake, R., Portele, C. & Whiteside, A. (eds). (2004). Geography Markup Language (GML), ISO Standard ISO/TC 211/WG 4/PT 19136 N005r3, OpenGIS Consortium.

Athanasiadis, I.N. (2007). Towards a virtual enterprise architecture for the environmental sector. In Protogeros, N. (ed.) Agent and Web Service Technologies in Virtual Enterprises. Chapter 15. Idea Group Inc.

Medeiros C.B., Perez-Alcazar J., Digiampietri L., Pastorello Jr. G.Z., Santanche A., Torres R.S., Madeira E., Bacarin E. (2005) WOODSS and the Web: Annotating and reusing scientific workflows, *SIGMOD Record* 34 (3), pp. 18-2

Bowers, S., Ludäscher, B. (2005) Actor-oriented design of scientific workflows. *Lecture Notes in Computer Science*, 3716 LNCS, pp. 369-384

Lee S., Wang T.D., Hashmi N., Cummings M.P. (2007) Bio-STEER: A Semantic Web workflow tool for Grid computing in the life sciences, *Future Generation Computer Systems* 23 (3), pp. 497-509