In **Fourth Biennial Meeting of the Int'l Environmental Modelling and Software Society**, (Sànchez-Marrè, M., Béjar, J., Comas, J., Rizzoli, A. E., and Guariso, G., Eds.), Barcelona, Spain, July 2008.

Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

A European database for integrated assessment and modeling of agricultural systems

Sander Janssen[1,2], Erling Andersen[3], Ioannis N. Athanasiadis[4] and Martin K. van Ittersum[1]

[1] Wageningen University, Plant Production Systems Group, Wageningen, sander.janssen/ martin.vanittersum@wur.nl

[2] Wageningen University, Business Economics Group, Wageningen

[3] University of Copenhagen, Forest and Landscape Group, Copenhagen, eran@life.ku.dk

[4] Dalle Molle Institute for Artificial Intelligence (IDSIA), Lugano, ioannis@idsia.ch

Abstract: Integrated Assessment and Modelling (IAM) can be used to assess socio-economic and environmental indicators, which generally require the linkage of models from different domains. To integrate a set of different models for an IAM, the data required by each of the models as inputs from a range of data sources also needs to be consistently integrated. This paper describes the process of development of a database integrating different data sources for an IAM project, and the human factors involved in the process of reaching consensus across peers with clashing requirements and needs. We adopted a structured process using a shared ontology as a means to one integrated relational database serving a set of models of a highly multi-disciplinary nature. The relational database covers data on agricultural systems, e.g. soil, climate, farm, agricultural management and agricultural policy data. The integrated database has been coupled to a range of quantitative models. The database schema and the shared ontology are distinct products that can be reused for or extended by other IAM projects requiring a similar set of data. It is recommended for any IAM project in which several models are coupled to adopt an explicit, collaborative and iterative process to specify an adequate data structure for storing data used in the project. For such a process to succeed it has to focus on the relevant domain knowledge captured across the data sources and this paper offers a proposal for such a process.

Keywords: Community modeling, agricultural systems, database, European Union

1. INTRODUCTION

Integrated Assessment and Modelling (IAM) is increasingly used to assess the impacts of policies, technologies or societal trends on the environmental, economic and social sustainability of systems (Parker, et al., 2002). IAM is a methodology to combine several quantitative models representing different systems and scales into a framework for Integrated Assessment (Parker, et al., 2002). Consequently, IAM can cover several organisational and spatio-temporal scales to provide quantitative assessment of impacts. To integrate a set of different models for an Integrated Assessment and Modelling project, the data required by each of the models as inputs and produced as outputs generally need to be consistently integrated. Each of the quantitative models used in an IAM is derived from a different discipline, requires different and to some extent overlapping data-sources, and is operational on different spatial and temporal scales.

SEAMLESS is an IAM research project (Van Ittersum, et al., 2008), which aims to provide a computerized framework to assess the sustainability of agricultural systems in the European Union at multiple scales. This aim is achieved by combining micro and marco level analysis, addressing economic, environmental and social issues, and facilitating the re-use of models and providing methods to conceptually and technically link different models together (Van Ittersum, et al., 2008).

Within SEAMLESS we faced a difficult data-integration challenge. Data have to serve dynamic biophysical models, static bio-economic farm models and partial computable general equilibrium market models. This required the integration multiple data-sources (including data related to European agriculture, including economic, biophysical, climatic data, model simulation input and output data, scientific workflow configurations and visualization of indicators) into a single relational database schema.

The objective of this paper is to describe the process of development of the SEAMLESS database, and the human factors involved in the process of reaching consensus across peers with clashing requirements and needs. The SEAMLESS European database on agricultural systems is presented. We adopted a structured process using a shared ontology as a means to arrive at one integrated relational database serving a set of models of a highly multi-disciplinary nature. This process is re-usable for other IAM projects, whereas the end result in terms of the database is re-usable for IAM of agricultural systems in Europe.

The next Section will describe firstly some theory behind ontologies and process of ontology engineering and the data sources of relevance to the SEAMLESS project. Consequently the results will be presented in the third Section as a description of the European database on agricultural systems, as the links between ontology and database and as the process used to construct this database with a group of researchers. Finally, conclusions and recommendations are provided.

2. MATERIAL AND METHODS

2.1 Ontologies and relational databases

In the context of integrated modelling, ontologies are useful to define the shared conceptualization of a problem, as ontologies consist of a finite list of concepts and the relationships between these concepts (Antoniou and van Harmelen, 2004) and as ontologies are written in a language, e.g. Web Ontology Language (McGuinness and van Harmelen, 2004), that is understandable by computers. In research aiming to integrate different models, scientists from various disciplines can define a common ontology that their domains share. A common ontology serves as a knowledge-level specification of the joint conceptualization, in our case of the data-sources used in the Integrated Assessment and Modelling project. Our efforts focused on the development of such a high-level ontology for the SEAMLESS data.

The common ontology is subsequently transcribed to the integrated relational database scheme, based on the conventions of the Semantic-Rich Development Architecture (SeRiDA) (Athanasiadis, et al., 2007). The SeRiDA combines object-oriented programming, relational databases and ontologies as three separate layers each with a distinct role: OWL ontologies for expressing rich domain semantics, Enterprise Java Beans™ for end-user application development, and normalized relational databases for persistence storage (Athanasiadis, et al., 2007). Through the SeRiDA the mapping of object-oriented models to ontologies is facilitated, while it provides an Object Relational Mapping (ORM), thereby acting as a bridge between different programming paradigms.

The use of ontologies has as advantages that the ontologies are richer in their representation of relationships between concepts than relational database schemas, have a strong implementation of inheritance, can be used as documentation tool for metadata, can be used for source code generation and allow to capture knowledge on the system under study as a distinct product.

2.2 Process of ontology engineering

In developing a common ontology, a group of scientists should agree and adopt one tight, well-reasoned and shared conceptualization. The development of a common ontology by a group of researchers is a complex, challenging and time-consuming task (Gruber, 1993, Holsapple and Joshi, 2002). Tools are available that help in ontology development and to store the ontology once it has been developed (e.g. Protégé OWL) (Knublauch, 2005). In developing the common ontology for the different data sources in our project, a collaborative approach was used. A collaborative approach is based on ‘development as a joint effort reflecting experiences and viewpoints of persons who intentionally cooperate to produce it’ and it thus requires a consensus-building mechanism (Holsapple and Joshi, 2002)’. As part of this collaborative approach, an inductive approach was used (Holsapple and Joshi, 2002). In our inductive approach, the common ontology was developed by

examining and analyzing the data-structures of the initial data-sources and extracting relevant properties or discussing the relationships.

2.3 Data sources

The data sources of relevance to the model-based assessments in the SEAMLESS project are:

(i) The Farm Accountancy Data Network (FADN) (EC, 2008a) is a source for evaluating the activities and income of agricultural holdings and the impacts of the Common Agricultural Policy. It consists of an annual survey carried out by the Member States of the European Union. The member states in the Union collect every year accountancy data from a sample of the agricultural holdings in the European Union (EC, 2008a). The data collected are, for example, physical and structural data, such as location, crop areas, livestock numbers, labor force, and economic and financial data, such as the value of production of the different crops, sales and purchases, production costs, production quotas and subsidies.

(ii) The European Soil Database (ESBN, 2008) on soils in Europe aims to provide a harmonised set of soil parameters, covering Europe (the enlarged EU) and bordering Mediterranean countries, to be used in agro-meteorological and environmental modelling at regional, national, and/or continental levels. Its scale is 1: 1.000.000 and it contains Soil Geographical Database of Eurasia, PedoTransfer Rules Database, Soil Profile Analytical Database of Europa and Database of Hydraulic Properties of European Soils (ESBN, 2008).

(iii) The European Interpolated Climate Data (JRC, 2008) provides interpolated daily data for a grid of 50 x 50 km covering Europe and Maghreb (average period 1975 -today). The majority of the original observations data originates from around 1500 meteorological stations across the European continent, Maghreb countries and Turkey.

(iv) Farm management data have been collected through dedicated surveys as part of the SEAMLESS project (Borkowski, et al., 2007). In the SEAMLESS project a lack of European data on agricultural management was identified. With agricultural management data is meant the use of inputs (fertilizers, pesticides, irrigation) and the timing of input use on farms. Surveys (Borkowski, et al., 2007) were developed as part of the SEAMLESS project. Data collected in these surveys are timing and amounts of inputs, crop rotations, machinery, labour requirements and costs.

(v) The COCO/CAPREG dataset (Britz, et al., 2007) is based on NewCronos (Eurostat, 2008) and FAOSTAT (FAO, 2008). It contains complete and mutually consistent time series for hectares/herd size, output coefficients, production, market balances, economic accounts and unit value prices (incl. consumer prices). For SEAMLESS, the relevant part of the COCO/CAPREG is the data on agricultural policies in the European Union.

The datasets from the Farm Accountancy Data Network, European Soil Database and European interpolated Climate data have been categorised into typologies (Metzger, et al., 2005, Andersen, et al., 2007, Hazeu, et al., 2007) to enable modelling of homogenous spatial units and to allow for characterization and sampling. The data sources haven been aligned with existing administrative categorizations like the Nomenclature of Territorial Units for Statistics (NUTS) (EC, 2008b).

3. RESULTS

3.1 Method to develop the integrated database

Initially, the data from the different sources were stored in eight different databases. To develop a common ontology for all data-sources, three scientists (a computer scientist, a landscape and forest ecologist and database expert, and a systems analyst) engaged in an integration process. These three scientists involved other domain experts in the integration process, when additional knowledge was required.

As a kick-off, a three day meeting was organized with experts on the database content and database set-up. Data-modeling was used to create a data-schema during the meeting. The result of this meeting was a database schema for some of the databases, which was subsequently translated into an ontology using Protégé (Knublauch, 2005). Next step was to extend this ontology by including all the relevant data sources required for running the models. This process lasted for over a period of six months with frequent discussions through email and web-meetings. During this period two additional face-to-face meetings were required of only one day. This first version of the common ontology was exported to

the first version of the SEAMLESS relational database schema using the SeRiDa-framework. The SEAMLESS database schema v.1.0 was discussed and improved between the three scientists involved in the project in roughly 3 iterations, leading to a first stable version of the database schema. Subsequently, the data from the original sources were entered into the database, which led to new improvements of the ontology and database schema v2.0.

When this database schema v2.0 was filled with data, the models were coupled to it using the Enterprise Java Beans™ generated by the SeRiDA framework. In coupling and running the models, some errors and required extensions of the common ontology were identified. These errors and required extensions were discussed and solved as part of the review of the database schema v2.0. During the review the three scientists tried to simplify and improve the schema as much as possible. This review lasted about two months and was organised through web-meetings and phone calls. Other domain experts were involved for their opinion on parts of the schema, which led to database schema v3.0. The data could be entered without requiring revisions into this version of the database schema. As part of the fourth version of the database schema, metadata will be included as part of the ontology.

3.2 European database on agricultural systems

Figure 1 provides an overview of the ontology developed for the European database on agricultural systems as developed in the SEAMLESS project. As can be seen from Figure 1, which shows the part of the database of relevance to soil, farm and climate data, there are concepts which classify the data, for example Farm Specialization, Farm Size and NUTS region and there are concepts that hold the actual data, like Representative Farm, Soil Characteristics and Daily Climate.

A central concept of the ontology is the concept of Representative Farm, which defines a FarmType in an FADN region in Europe for a specific year. A FarmType is specified according to the dimensions of farm size, farm intensity and farm specialization (Andersen, et al., 2007) (Fig. 2). As an example of a classifying concept, Farm intensity is a classification of farms according to their total output of agricultural produce per hectare (Andersen, et al., 2007). If the total output is below 500 euros per hectare, then the farm falls in the class of low intensity, if it is between 500 and 3000 euros, then it is medium intensity and if is more than 3000 euros, then it is high intensity. While a FarmType is not linked to a specific region or year, a Representative Farm is specific to a region and a year.

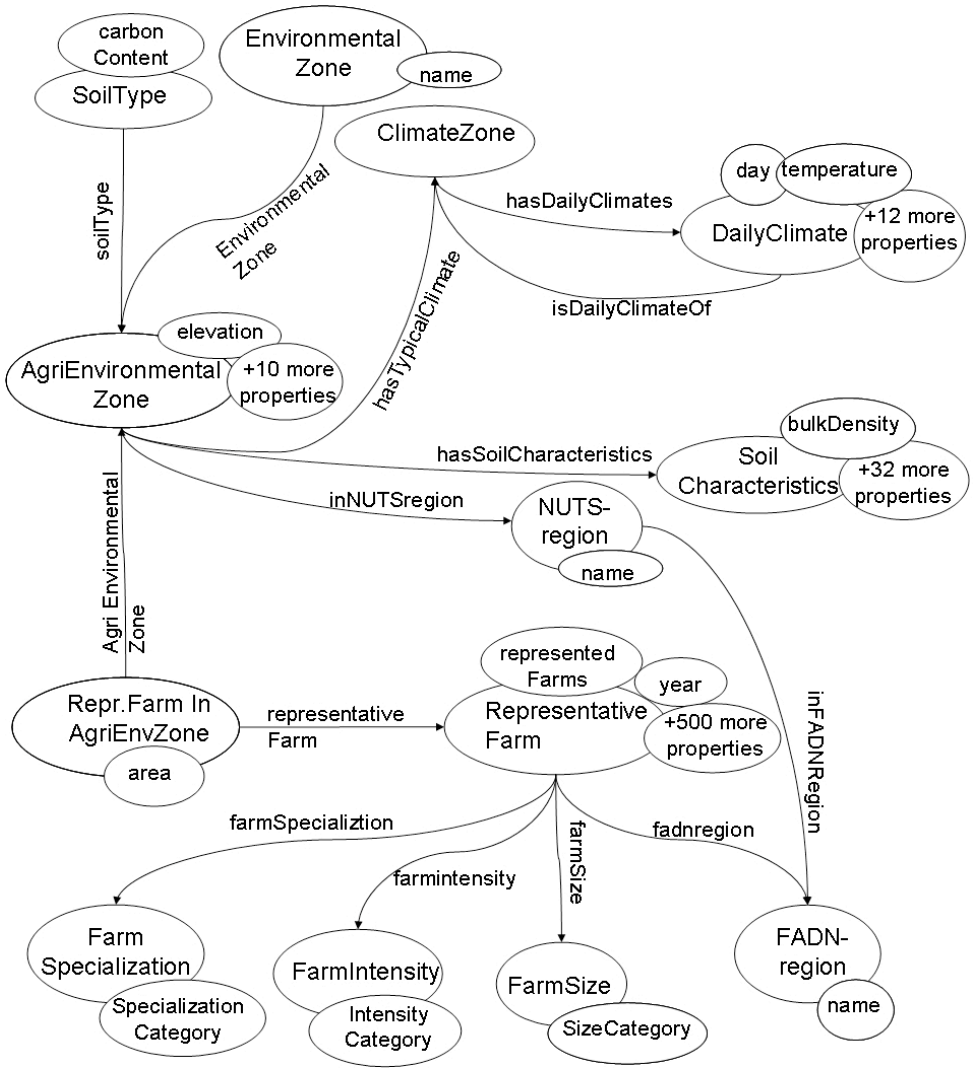

Figure 1. An ontology-schema of the European database on agricultural systems showing the parts on farms, soils, climate and their links.

As can be seen in Figure 1, AgriEnvironmental Zone is a central concept, in that it links to soil and climate data. An AgriEnvironmental Zone is a unique combination of an Environmental Zone, the soilType and NUTS region. AgriEnvironmental Zones are the smallest homogenous units in a region in terms of climate and soil data. Environmental zones are used to stratify the diverse European Union climate in zones with a similar climate (Metzger, et al., 2005). The Environmental Zones cover more than one region, and a Climate Zone is thus a unique combination for a NUTS-2 region and Environmental Zone for which a set of climate data is available. A Climate Zone provides the daily climate data for a 30-years time period for a region and Environmental Zone, so one record for every day. Examples of properties of daily climate data are rainfall in mm per day, average daily temperature in degrees Celsius per day and wind speed at 10m in m/s.

Each AgriEnvironmental zone is linked to a set of soil data, as classified according to Soil Types. Six different Soil Types were defined according to topsoil organic carbon classes (Hazeu, et al., 2007). For each unique combination of a Soil Type and a NUTS-region a set of soil data is available as stored in the concept of Soil Characteristics. Examples of properties of the soil characteristics are thickness subsoil and topsoil, depth to rocks and saturation top soil.

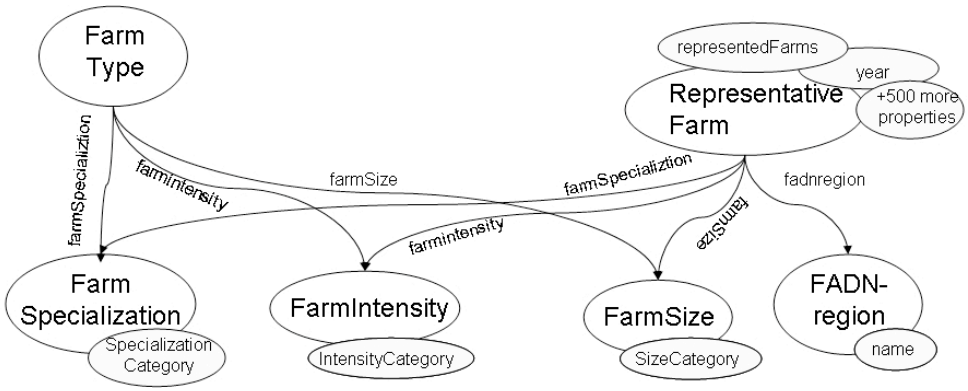

Figure 2. The concepts Farm Type and Representative Farm and the relationships to their classifying concepts.

The link between AgriEnvironmental Zones and Representative Farms is made through allocating an area of an AgriEnvironmental Zone to each Representative Farm. This implies that each AgriEnvironmental Zone is allocated to one or more Representative Farms and each Representative Farm can be found in one or more AgriEnvironmental Zones. As can be seen from Figure 1, Representative Farms and AgriEnvironmental Zones are based on different administrative regions e.g. AgriEnvironmental Zones refer to NUTS-2 regions (EU25 has 270 NUTS-2 regions) and Representative Farms refer to FADN-regions.

In this paper the link between agricultural management data, policy data, AgriEnvironmental Zones and Representative Farms will not be explained in detail. As the agricultural management differs within regions, Regional Agricultural Management Zones were created. A Regional Agricultural Management Zone has a distinct set of agricultural management data and is linked to one or more Agri-Environmental Zones. Finally, data on agricultural policies and prices are linked to each NUTS-region. The current version of the database consists of 329 tables including 2 035 fields and with 379 relations between the tables. The number of records in the database now exceeds 7.4 million.

.3 Links between ontology and relational database

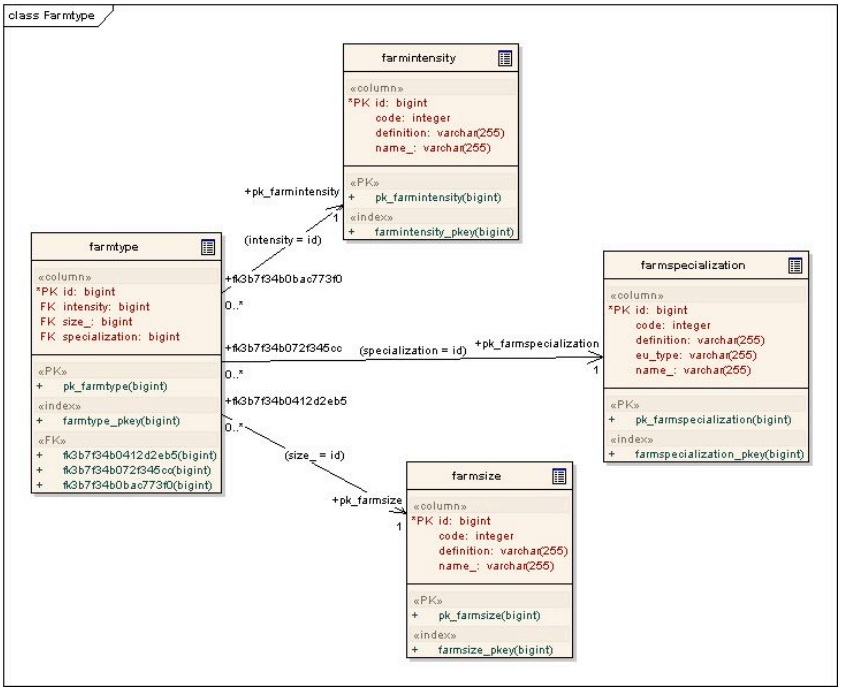

Figure 3. a relational database schema of the FarmType, FarmIntensity, FarmSize and FarmSpecialization-tables (made with Enterprise Architect©).

Figure 3 presents part of relational database schema related to FarmType as it is generated from the ontology schema from Figure 2. From Figure 3 it can be seen that all the relationships between the tables are enforced through foreign and primary keys. The

FarmType table is linked by many-to-one foreign keys to the classifying tables FarmIntensity, FarmSize and FarmSpecialization. These many-to-one foreign keys represent the relationships farmSpecialization, farmIntensity and farmSize from Fig. 2, which describe that each FarmType has one and only one reference to the classifying concepts of FarmIntensity, FarmSize and FarmSpecialization. This example demonstrates the translation of the ontology into the relational database schema that is usable for persistent data storage. More examples can be found in Athanasiadis, et al.(2007).

4. CONCLUSION AND RECOMMENDATIONS

By using ontologies in a collaborative process of conceptual modelling, we managed to derive a common database schema that integrates a range of data sources from different domains specified at different spatial and temporal scales. This common database schema and the common ontology on which it is based are distinct products that can be reused for or extended by other research projects requiring a similar set of data. The integrated database has been linked to a range of quantitative models and can be coupled to other models with similar data requirements. It is recommended for any Integrated Assessment and Modelling project in which several models are linked or complex models are developed to adopt an explicit process to specify an adequate data structure for storing data used in the project. This paper provides a proposal for such a process, which should be collaborative and iterative. Using a framework like SeRiDA for mapping between programming paradigms allows the programmers to benefit from the strengths of each of the programming paradigms. Also, adopting an explicit process to specify an adequate data structure and a framework likes SeRiDA helps scientists to focus on the domain content of the data structure, while not loosing focus in details of technical implementation in different programming paradigms.

ACKNOWLEDGMENTS

We thank all scientists in the SEAMLESS project who contributed to development of the common ontology on projects and scenarios. This work has been carried out as part of the SEAMLESS Integrated Project, EU sixth Framework Programme, Contract No. 010036-2.

REFERENCES

Andersen, E., B. Elbersen, F. Godeschalk and D. Verhoog, Farm management indicators and farm typologies as a basis for assessments in a changing policy environment, *Journal of Environmental Management*, 82(3), 353, 2007.

Antoniou, G. and F. van Harmelen, *A Semantic Web Primer*, The MIT Press, 238 pp., Cambridge, Massachusetts; London, England, 2004.

Athanasiadis, I. N., F. Villa and A. E. Rizzoli, Enabling knowledge-based software engineering through semantic-object-relational mappings, paper presented at 3rd International Workshop on Semantic Web Enabled Software Engineering, 4th European Semantic Web Conference, Innsbruck, Austria, 2007.

Borkowski, N., P. Zander, G. Stokstad, Sandra Uthes, F.-J. Reinhardt, G. Bellocchi and M. Hecker, Conceptual Approach to Identify and Assess Current Activities, PD3.3.9, SEAMLESS integrated project, EU 6th Framework Programme, contract no. 010036-2, www.SEAMLESS-IP.org, 23 pp. pp., Wageningen, 2007.

Britz, W., I. Pérez, A. Zimmermann and T. Heckelei, Definition of the CAPRI Core Modelling System and Interfaces with other Components of SEAMLESS-IF, SEAMLESS Report No.26, SEAMLESS integrated project, EU 6th Framework Programme, contract no.010036-2, 116 pp., 2007.

EC, Farm Accountancy Data Network (FADN) Source: EU-FADN-DG AGRI-G3, European Commission. Retrieved on 15 February 2008 from http://ec.europa.eu/agriculture/rica/.

EC, Nomenclature of territorial units for statistics - NUTS Statistical Regions of Europe, European Commission. Retrieved on 15 February 2008 from http://ec.europa.eu/comm/eurostat/ramon/nuts/home_regions_en.html.

ESBN, European Soil Database, European Soil Bureau Network, European Commission-Joint Research Centre (JRC). Retrieved on 15 February 2008 from http://eusoils.jrc.it/esbn/Esbn_overview.html.

Eurostat, European Dissemination Database (Formerly New Cronos), Retrieved on 15 February 2008 from http://epp.eurostat.ec.europa.eu/.

FAO, FAOSTAT, Food and Agriculture Organisation of the United Nations. Retrieved on 15 February 2008 from http://faostat.fao.org/.

Gruber, T. R., A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition*, 5, 199-220, 1993.

Hazeu, G., B. Elbersen, E. Andersen, B. Baruth, C. A. van Diepen and M. J. Metzger, The SEAMLESS biophysical typology: a spatial agri-environmental modeling framework, In: Brouwer, F. and M. K. van Ittersum (Ed.) *Unknown; forthcoming*, Springer Academic Publishing, 2007.

Holsapple, C. W. and K. D. Joshi, A collaborative approach to ontology design, *Communications of the ACM*, 45(2), 42-47, 2002.

JRC, Meteorological data Source JRC/AGRIFISH Data Base - EC - JRC, AGRIFISH unit-Institute for the Protection and Security of the Citizen (IPSC) European Commission - Joint Research Centre. Retrieved on 15 February 2008 from http://agrifish.jrc.it/marsstat/datadistribution/.

Knublauch, H., Protege OWL, Stanford Medical Informatics. Retrieved on 24 January 2008 from http://protege.stanford.edu/.

McGuinness, D. and F. van Harmelen, OWL Web Ontology Language Overview, WWW Consortium. Retrieved on 24 January 2008 from www.w3.org/TR/owl-features/.

Metzger, M. J., R. G. H. Bunce, R. H. G. Jongman, C. A. Mucher and J. W. Watkins, A climatic stratification of the environment of Europe, *Global Ecology and Biogeography*, 14(6), 549-563, 2005.

Parker, P., R. Letcher, A. Jakeman, M. B. Beck, G. Harris, R. M. Argent, M. Hare, C. Pahl-Wostl, A. Voinov, M. Janssen, P. Sullivan, M. Scoccimarro, A. Friend, M. Sonnenshein, D. Barker, L. Matejicek, D. Odulaja, P. Deadman, K. Lim, G. Larocque, P. Tarikhi, C. Fletcher, A. Put, T. Maxwell, A. Charles, H. Breeze, N. Nakatani, S. Mudgal, W. Naito, O. Osidele, I. Eriksson, U. Kautsky, E. Kautsky, B. Naeslund, L. Kumblad, R. Park, S. Maltagliati, P. Girardin, A. Rizzoli, D. Mauriello, R. Hoch, D. Pelletier, J. Reilly, R. Olafsdottir and S. Bin, Progress in integrated assessment and modelling, *Environmental Modelling & Software*, 17(3), 209, 2002.

Van Ittersum, M. K., F. Ewert, T. Heckelei, J. Wery, J. Alkan Olsson, E. Andersen, I. Bezlepkina, F. Brouwer, M. Donatelli, G. Flichman, L. Olsson, A. Rizzoli, T. van der Wal, J.-E. Wien and J. Wolf, Integrated assessment of agricultural systems- a component based framework for the European Union (SEAMLESS), *Agricultural Systems*, 96, 150-165, 2008.