The use of UML to design agricultural data warehouses

François Pinet[1], André Miralles[2], Sandro Bimonte[1], Françoise Vernier[3], Nadia Carluer[4], Véronique Gouy[4], Stephan Bernard[1]
[1] Cemagref, 24 avenue des Landais, 63172 Aubière
[2] Cemagref, 361, rue Jean-François-Breton, 34196 Montpellier
[3] Cemagref, 50 avenue de Verdun, 33612 Cestas
[4] Cemagref, 3 Bd Quai Chauveau, 69009 Lyon

Abstract
Recent research works propose to use the Unified Modeling Language (UML) to design data warehouses. First, the paper overviews these recent UML-based techniques. We show that UML can help system designers to build a data warehouse model. This type of model aims to describe the different analysis dimensions of the data. Second, we will also present different UML-based tools used during a project of agricultural data warehouse. The paper will introduce this data warehouse developed in France. Its goal is to allow analyzing spatially the use of pesticide in agriculture.

1. Introduction

Along with the development of new information and communication technologies, we have seen an increase in sources of agricultural data. For example, more and more systems use sensors and satellite images to monitor work being done in agriculture. Farmers enter other data with specialized computer programs (application that records agricultural practices). In order to analyze the agricultural data issued from different sources with environmental information, it is needed to use tools and methods to integrate these information. Data warehouses are a specific type of database that serves to integrate, accumulate and analyze data from various sources (Calì et al. 2003). Information stored in different databases can be group together in a data warehouse for combined analysis. Depending on their requirements, one can load data every week, every month, every year or even less frequently. These data are usually organized in a form that speeds up calculation of indicators. The indicators are made up of aggregated information obtained by aggregation functions such as sum, average, variance, etc. Using data warehouses is therefore important within a decision-making context. For example, a data warehouse containing economic, urban and environmental information will help decision-makers find the best place to establish a new infrastructure. The concept of the data warehouse has great potential for assessing the impact of actions, practices, scenarios and programs from both the socio-economic and the environmental point of view (Schneider 2008; Nilakanta et al. 2008; Schulze et al. 2007; Mahboubi et al., 2010; Bimonte, 2010).

Recent research works propose to use the Unified Modeling Language (UML) to design data warehouses. First, the paper will overview these recent UML-based techniques. We will show that UML can help system designers to build a data warehouse model. This type of model aims to describe the different analysis dimensions of the data (Malinowski and Zimanyi 2008). Second, we will also present different UML-based tools used during a project of agricultural data warehouse. The paper will introduce a first prototype of this data warehouse developed in France. Its goal is to allow analyzing spatially the use of pesticide in agriculture.

2. Data warehouses: main concepts

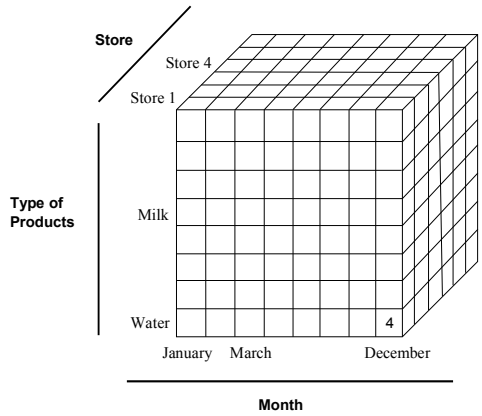
In this section, we present an example from Trujillo et al. (2001). The ***facts*** of a data warehouse are the values of the indicator to analyze (Malinowski and Zimanyi 2008). In the example, we consider the facts of the data warehouse to be the product sales of a company in dollars. Each of the company's stores provides these data. In a data warehouse, an analysis results from the use of an aggregation operation (e.g., sum or average) on the facts. In the example, a possible analysis is the sum of sales calculated by category of product, by store and by month. The result of this analysis can be represented in a cube (Trujillo et al. 2001) - see Figure 1.a. Each dimension of the cube corresponds to a criterion of analysis: type of products, store and month. The cells of the cube are called ***measures***. They store the sums of sales for each tuple <type of products, store, month>. For instance, in Figure 1.a, the sum of sales for the tuple <Water, Store 1, December> is 4. In data warehouses, the criteria of analysis are structured in hierarchies called ***dimensions***. Figure 1.b shows the three dimensions presented by Trujillo et al. (2001). A data warehouse can produce many analyses by combining different levels of dimensions. For example, other cubes could be calculated:

- sums of sales by city,
- sums of sales by brand, by city, by year,
- sums of sales by type, by state, by season, etc.

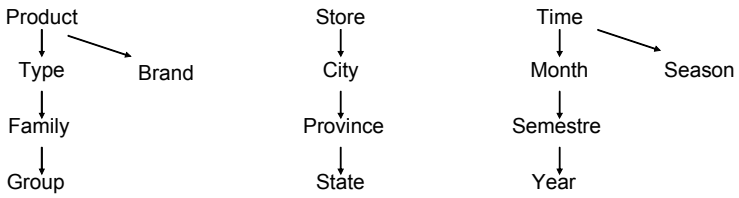
Note that data warehouses generally support *n*-dimensional cubes. Data can be combined to provide previously unknown causal links. To do so, users can visualize cubes from the data warehouse using tools like OLAP (On-line Analytical Processing) (Malinowski and Zimanyi 2006; Malinowski and Zimanyi 2008). Causal links can also be discerned automatically with data-mining algorithms (Berson and Smith 1997).

Using data warehouses is therefore important within a decision-making context. For example, a data warehouse containing economic, urban and environmental information will help decision-makers find the best place to establish a new infrastructure. The concept of the data warehouse has great potential for assessing the impact of actions, practices, scenarios and programs from both the socio-economic and the environmental point of view (Schneider 2008). Two examples of use in agriculture can be found in (Nilakanta et al. 2008; Schulze et al. 2007).

Multi-dimensional models aim to describe the facts and the different analysis dimensions of a data warehouse (Malinowski and Zimanyi 2008). Recently, some articles have presented specific methods for formalizing multi-dimensional models.



a) Cube storing sales by category of products, by store and by month



b) Analysis dimensions

Figure 1. Example of a data warehouse

3. UML for data-warehouse modelling

Several authors have proposed object oriented multi-dimensional models, some of them being based on an extension of UML. The models of (Herden 2000; Nguyen et al. 2000; Prat et al. 2006; Trujillo et al. 1998) are basic models which incorporate and formalize essential notions (fact classes, dimension classes, roll-up associations, measures) and offer a number of specific features.

The work of (Trujillo et al. 1998) introduces the notion of cube classes with a set of possible operations to define the analysis. In (Herden 2000), a fact class can be aggregated from several others fact classes. A UML profile is suggested. It can be manipulated through an extension of the CASE Tool Rational Rose. The models of (Luján-Mora et al. 2006) are much more sophisticated. In (Abelló et al. 2006), there are 6 types of nodes in the multidimensional graph. Various types of associations are available. It is possible to change the dimensions of a cube. In (Luján-Mora et al. 2006), typical multidimensional structures are defined through packages. 14 stereotypes are suggested for packages, classes, associations and attributes. Each of these two models is supported by a specific UML profile. Since this profile is complex to manage, its correct use is controlled through the definition of constraints in natural language or in expressions in Object Constraint Language (OCL). Concerning the design and the implementation of multidimensional systems, various propositions have been made depending on a given context or platform. In (Hahn 2000) an environment is suggested

which is able to generate the implementation of a star or snowflake multidimensional structure from a conceptual schema. The generation process takes into account the limitations of the OLAP target system (Cognos Powerplay or Informix Metacube). The work of (Moody et al. 2000) also suggests a method for developing dimensional models from E/R schemas. Different options for the resulting schema can be chosen (flat, star, snowflake, constellation). In (Prat et al. 2006), a multidimensional structure is derived from a UML schema. The work of (Theodoratos et al. 2008) addresses the problem of integrating the data from heterogeneous databases and storing it in the repository of the multidimensional structure. In this work, the multidimensional structure is seen as a set of materialized views. So the problem becomes one of view selection. Different algorithms are proposed and compared for solving it. The work of (Mazon and Trujillo 2008) describes how to align the whole data warehouse development process with an Model Driven Architecture framework.

Some proposals of formalization have been introduced to model spatial data warehouse (Malinowski and Zimanyi 2008; Pestana et al., 2005). Several uses of spatial data warehouse are evocated, for example in (Bernier et al., 2009, Julien et al., 2009, McHugh et al., 2009, Sboui et al., 2010; Bimonte, 2010). This type of data warehouse is important in agriculture and environment; environmental data are very often georeferenced.

4. Example of data warehouse for the analysis of the pesticides

In the context of a French project, several data sources about the pesticide use in agriculture have been integrated in a spatial data warehouse. It enables visualising the different types of pesticides used over the time as well as their spatial distribution in a watershed. This data warehouse allows calculating the total quantities of pesticides applied on the cultivated parcel during at different levels of granularity.

We model the spatial data warehouse in UML, in using formalization close to the one proposed by (Pestana et al., 2005). This visual formalism was implemented as profile (technology similar to that of the pluggings) into the UML-based tool Objecteering developed by Softeam. This profile is called SOLAP Profile. In this visual formalism, the Pesticide Cube is annotated with the pictogram (Figure 4).

Figure 4. Global model of the Pesticide Cube

This global model is detailed into Figure 5. The parameters interesting the users are mentioned into the Measure, concept which is annotated with the pictogram  . In this prototype of cube, the parameter chosen is the quantity of Active Substance Spread on each parcel at each application. This parameter (i.e. the measure) is analyzed according to the three criteria:

1. the active substance used by the farmers represented in the model by the Active Substance Dimension; this dimension is a thematic dimension since it is annotated with the pictogram ;
2. the spatial organisation of the territory represented by the Spatial Dimension which is annotated with the pictogram ;
3. the temporal analysis represented by the Temporal Dimension annotated with the pictogram .

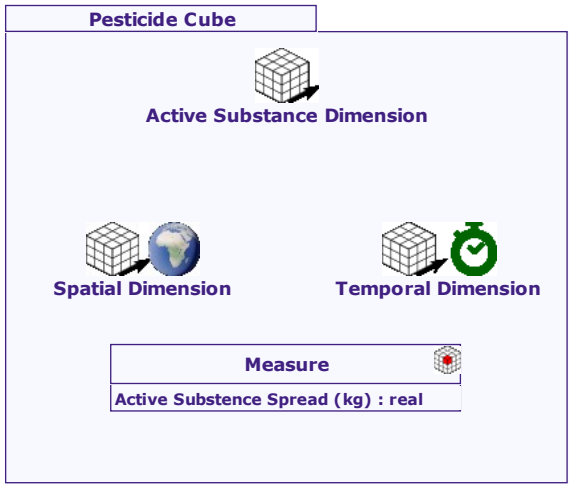


Figure 5. Detailed model of the Pesticide Cube

As described previously, the three criteria of analysis are organized according to different levels of granularity interesting the users or the managers. These levels of granularity are modelled for each dimension into Figures 6, 7 and 8.

Figure 6 shows that the finest level of granularity is Active Substance. The fine level of Active Substance is aggregated into Active Substance Type (herbicide, fungicide and insecticide) and, finally, at the top level All Active Substances are aggregated.

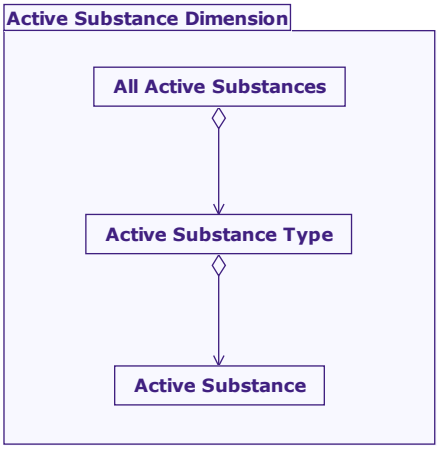


Figure 6. Model of the levels of granularity of the Active Substance Dimension

The Spatial Dimension (Figure 7) is very simple with its two levels: the fine level is represented by the cultivated Parcel and the upper level by the Catchment basin.

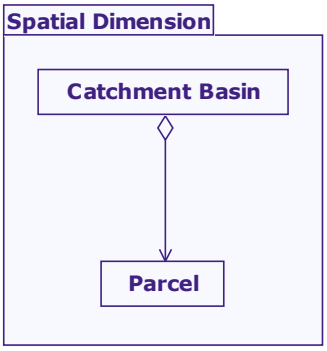


Figure 7. Model of the levels of granularity of the Spatail Dimension

The Temporal Dimension (Figure 8) has four levels of granularity: the fine level which records the day where the application is done (Application Day), the Week level which allows a the seasonal partially comparison, the Year level which gives the yearly evolution of the applied quantity of pesticide and, of course, at the top level, the total of the pesticides applied since the beginning of the records.

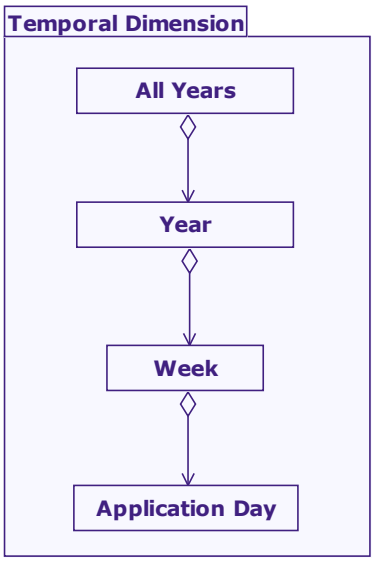


Figure 8. Model of the levels of granularity of the Temporal Dimension

Once the model of the cube done, it must be implemented. The physical structure of the data warehouse uses the database technology. A functionality was developed in the SOLAP Profile allowing to transform the model of a cube into the model of the database implementing the cube. Figure 9 shows the schema of the database of the Pesticide Cube.

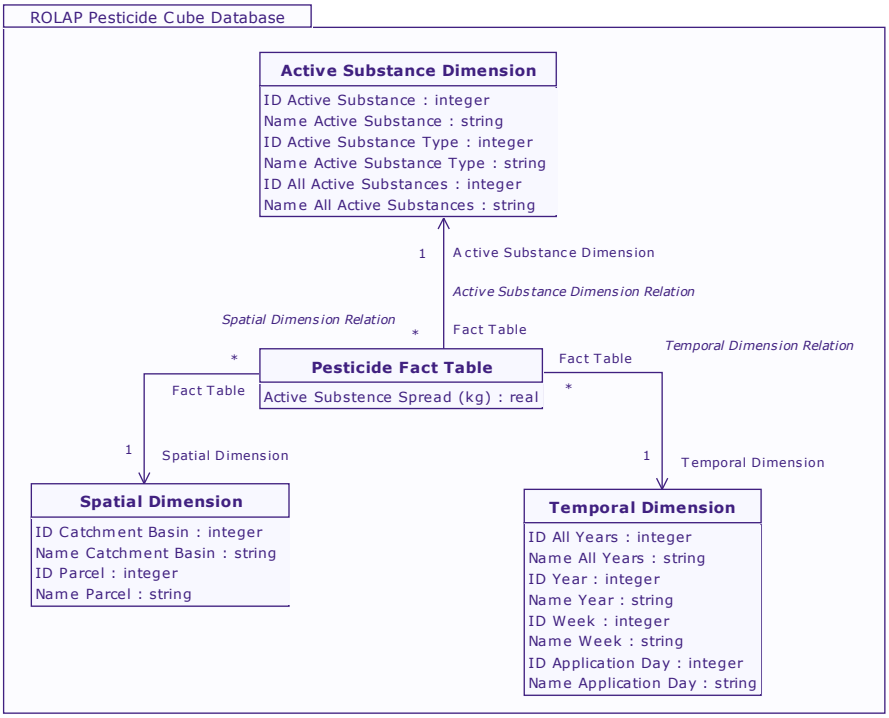


Figure 9. Schema of the Database implementing the cube for the pesticides

The database has been generated and the data sources have been integrated with the ETL tool (Extraction, Transformation, Loading) called Spatial Data Integrator developed by Talend. Users can interactively visualise the data through geographical maps thanks to the Map4Decesion tool implementing the Spatial On-Line Analytical. This tool is the result of the researches of Industrial Research Chair in Geospatial Databases for Decision Support of the Université Laval. Figure 10 sums up the different tools used to build and exploit a the data warehouse for the pesticides.

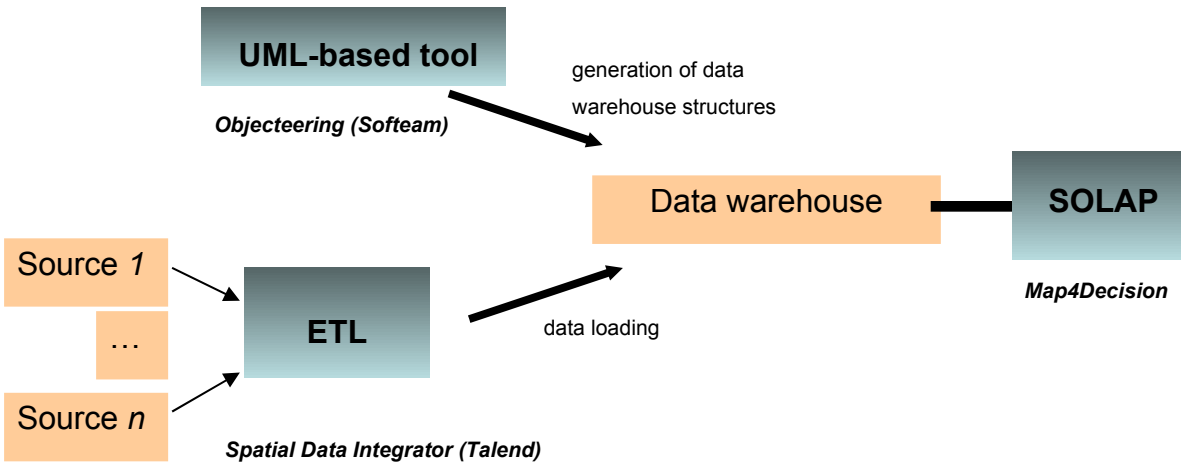


Figure 10. Tools involved in the pesticide data warehouse

Once the cube implemented and the data loaded, uses and managers can access the data stored in the cube to make different series of analysis coupling different levels of granularity of the three parameters modelized,by the dimension. The Map4Decision's screenshot of

figure 11 gives an example of analysis. It shows the quantity of pesticides applied on the parcels two successive years. Visually, a manager can see that the quantity of pesticide applied the second year is more important than the first. Moreover, one can apply SOLAP operators to navigate into the spatial dimension by the simple interaction with maps. In this way, the decision-makers can explore the spatial data warehouse in an easy way, discovering unknown relationships and trends concerning pesticides in France.

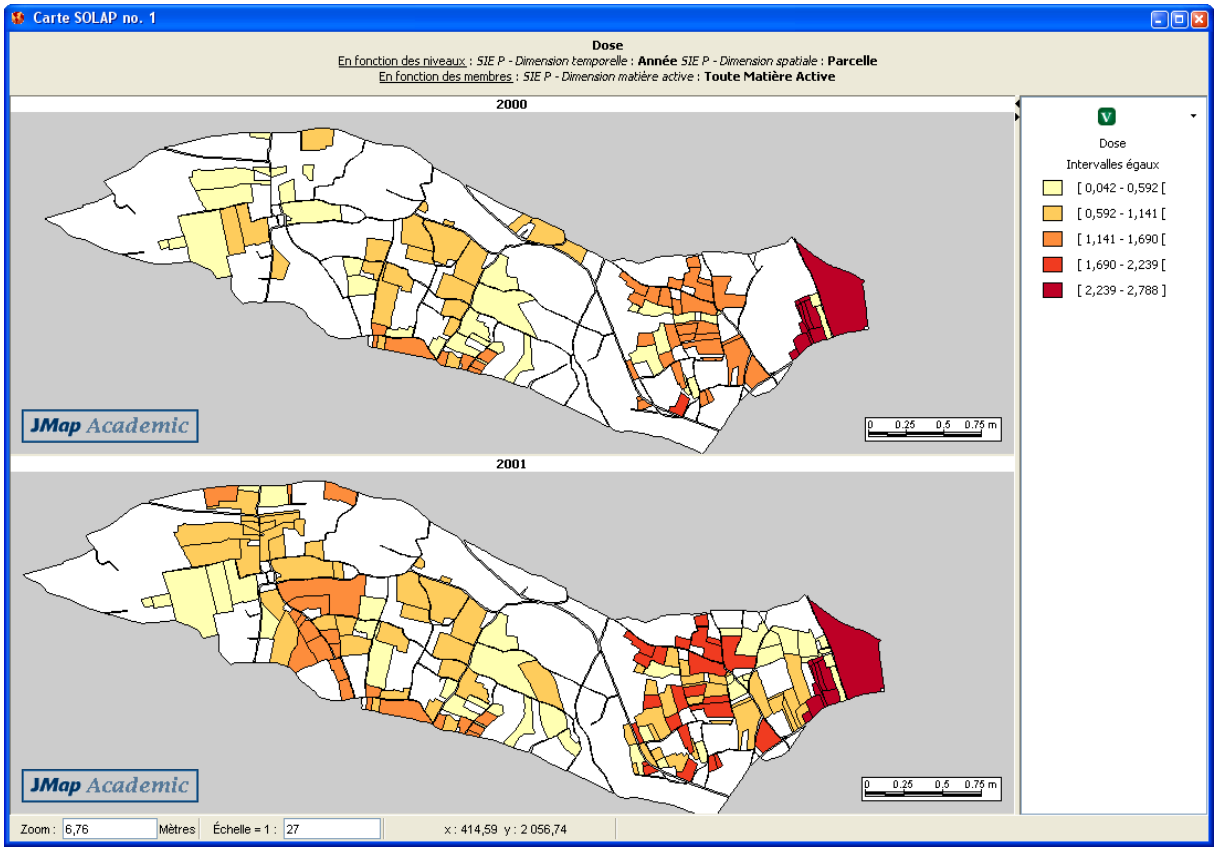


Figure 11. Temporal evolution of total pesticide applied on the parcels of a catchment basin.

5. Conclusion and example of design of an agricultural data warehouse

People use data warehouses to help them make decisions. For example, public policy decision-makers can improve their decisions by using this technology to analyze the environmental effects of human activity. In production systems, data warehouses provide structures for extracting the knowledge required to optimize systems. While use of data warehouses is becoming widespread in certain spheres of activity such as finance and mass marketing, its use in environmental and agricultural fields is still marginal. This technology may, however, be of great service in these sectors, whether in decision-making or in optimization of agricultural and environmental systems.

Designing data warehouses is a complex task; designers need flexible and precise methods to help them create data warehouses and adapt their analysis criteria to developments in the

decision-making process. UML can be used to facilitate the modelling of data warehouses (see the previous section).

The use of UML in this project has facilitated the modelling of the data warehouse and the communication with the protagonists of the projects.

The prospects for research in the field of UML modelling of data warehouses are still numerous and varied; see Rizzi et al. (2006) for other examples of prospects. We think that UML will be of great service in the implementation of data warehouses.

References

Calì A., Lembo D., Lenzerini M., Rosati R. (2003) Source Integration for Data Warehousing. Multidimensional Databases, pp 361-392

Malinowski E., Zimanyi E. (2008) Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications, Springer, 435 p

Nilakanta S., Scheibe K., Rai A. (2008) Dimensional issues in agricultural data warehouse designs. Computers and Electronics in Agriculture vol. 60(2), pp 263-278

Rizzi S., Abello A., Lechtenborger J., Trujillo J. (2006) Research in data warehouse modeling and design: dead or alive? Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, pp 3-10

Schulze C., Spilke J., Lehner W. (2007) Data modeling for Precision Dairy Farming within the competitive field of operational and analytical tasks. Computers and Electronics in Agriculture vol. 59 (1), pp 39-55

Schneider. M. (2008) A general model for the design of data warehouses. International Journal of Production Economics vol. 112(1) pp 309-325

Herden O. (2000) A Design Methodology for Data Warehouses. In: Proc. of the CAISE Doctoral Consortium, Stockholm

Nguyen T.B., Tjoa A.M., Wagner R. (2000) An Object Oriented Multidimensional Data Model for OLAP. In: Lu, H., Zhou, A. (eds.) WAIM 2000. LNCS, vol. 1846, pp. 69–82. Springer, Heidelberg

Prat N., Akoka A., Comyn-Wattiau I. (2006) A UML-based data warehouse design method. Decision Support Systems (DSS) 42(3), 1449–1473

Trujillo J., Palomar M. (1998) An object Oriented Approach to Multidimensional Database Conceptual Modeling (OOMD). In: Proc. of the 1st ACM international workshop on Data warehousing and OLAP, Washington, 16–21

Abelló A., Samos J., Saltor F. (2006) YAM2: A Multidimensional Conceptual Model Extending UML. Information Systems 31, 541–567

Luján-Mora S., Trujillo J., Song I.Y. (2006) A UML Profile for Multidimensional Modeling in Data Warehouses. Data & Knowledge Engineering 59, 725–769

Hahn K., Sapia C., Blaschka M. (2000) Automatically Generating OLAP Schemata from Conceptual Graphical Models. In: Proc. of the 3rd ACM International Workshop on Data Warehousing and OLAP, DOLAP 2000, McLean, Virginia, USA

Moody L.D., Kortink M.A.R. (2000) From Enterprise Models to Dimensional Models: A Methodology for Multidimensional Structure and Data Mart Design. In: Proc. of the International Workshop on Design and Management of Multidimensional structures, DMDW 2000, Stockholm, Sweden

OMG: OCL 2.0 specification version 2.0. OMG specification, 185 pages (2005), http://www.omg.org

Mazon J.N., Trujillo J. (2008) An MDA Approach for the Development of Data Warehouses. Decision Support Systems (45), 41–58 (2008)

Theodoratos D., Ligoudistianos S., Sellis T. (2001) View selection for designing the global Multidimensional structure. Data & Knowledge Engineering 39, 219–240

Pestana G., Da Silva M. M., Bedard Y. (2005) « Spatial OLAP modeling: an overview base on spatial objects changing over time », Computational Cybernetics, 2005. ICCC, p.149-154.

Mahboubi H., Faure T., Bimonte S., Deffuant G., Chanet J.P., Pinet F. (2010) A Multidimensional Model for Data Warehouses of Simulation Results. To appear in: International Journal of Agricultural and Environmental Information Systems vol. 1 (2), IGI Global

Bernier E., Gosselin P., Badard T., Bédard Y., Easier Surveillance Of Climate-Related Health Vulnerabilities Through A Web-Based Spatial Olap Application, International Journal of Health Geographics, vol.8 (18) 2009.

Julien F. S., Rivest S., Les analyses relatives au transport maritime : un exemple de géodécisionnel, Colloque Géomatique 2009, 21-22 octobre, Montréal, Canada, 2009.

Mchugh R., Roche S., Bédard Y. (2009) Towards a SOLAP-based public participation GIS, Journal of Environmental Management, vol. 90(6) 2009, p.2041-2054.

Sboui T., Salehi M., Bédard Y. (2010) A systematic approach for managing the risk related to semantic interoperability between geospatial datacubes, *to appear in: International Journal of Agricultural and Environmental Information Systems*, vol. 1(2), 2010.

Bimonte S. (2010) A Web-Based Tool for Spatio-Multidimensional Analysis of Geographic and Complex Data, to appear in: International Journal of Agricultural and Environmental Information Systems, vol.1 (2), 2010.

Rizzi S., Abello A., Lechtenborger J., Trujillo J. (2006) Research in data warehouse modeling and design: dead or alive? Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, pp 3-10