Investigating a Bottom-up Approach for Extracting Ontologies from Urban Databases

Christophe Chaidron [1], Roland Billen [1] and Jacques Teller [2]

[1] Geomatics Unit, University of Liege,
17 Allée du 6-Août, B-4000 Liege, Belgium
{cchaidron,rbillen}@ulg.ac.be
[2] Fonds National de la Recherche Scientifique
LEMA Université de Liège, Lab. of Architectural Methodology
1 Chemin des Chevreuils, B52/3, 4000 Liège, Belgium
jacques.teller@ulg.ac.be

1 Introduction

Ontologies, as "formal and explicit specifications of shared conceptualizations" [1] play a predominant role when developing information systems. This role is increasingly recognised by geo and urban experts when dealing with urban (geo) spatial information systems (GIS, SIS) and spatial databases (SDB). Generally speaking, they provide significant benefits for the design and use of geographic information, such as defining semantics independently of data representation [2]. Urban GIS and SDB are therefore a large source of urban "domain" ontologies [3], like technical networks, urban planning concepts, cadastre structures etc.

It is not yet a common practice to record explicit formalisation of concepts in GIS-SDB documentations. The reason for this is that most GIS-SDB designers have no specific background in ontology design and the role or usefulness of ontologies is still largely underestimated by practitioners. It is hence quite common to have no trace of ontologies in current urban information systems. They are hidden behind documentation, files, database tables or simply part of implicit experts' knowledge. Extracting ontologies from such disparate sources is not a trivial task as it may reveal inconsistencies or gaps in the semantic model underlying these databases.

The aim of this paper is to investigate a *bottom-up approach* for extracting *local ontologies* from urban databases. By local ontologies we mean ontologies related to the databases themselves. Local ontologies of urban SDB contain information about urban phenomena and therefore could be used to (re)construct urban domain ontologies. Different ontology design methods have been presented in the literature, including bottom-up [4] and top-down [5] approaches. A more detailed presentation of such methods can be found in [3]. When dealing with a non-well documented GIS or SDB, this article suggests that starting with defining specifics notions and then extracting more generic concepts by generalisation appears as a pragmatic way to handle ontology generation (extraction).

The paper is organised as follows. First we remind SDB definition and roles of ontology in SDB design. Then we present the empirical bottom-up approach we

C. Chaidron et al.: *Investigating a Bottom-up Approach for Extracting Ontologies from Urban Databases,* Studies in Computational Intelligence (SCI) **61**, 131–141 (2007)
www.springerlink.com

recommended and the next section presents the case study where the approach has been adopted. Finally we draw short conclusions.

2 Spatial databases

There are various sources of information about cities and urban phenomena (plans, maps, registers, etc.). Nowadays, most of the information is stored in numerical format; especially, geographical (spatial) information about urban areas is mostly stored in SDB or GIS. The specificity of these databases is their capacity of storing spatial data, i.e. geographical entities that are described by attributes (standard tuples of a database: alphanumerical data or images, sounds, binary attributes…) associated to some geometric information (position, shape, geometrical and topological relationships, etc…).

Despite an extensive diffusion of such spatial systems and their common use by many citizens, especially through internet (navigation routing, location-based services, visualisation such as "Google Earth", etc.), their conception is not within anybody's reach. System's designers have to follow a formalised methodology, laying stress on the modelling step. More particularly, it requires the creation of specialised documents, according to international standards, like feature catalogue, formalised conceptual data models, and using dedicated tools (Computer-Aided Software Engineering).

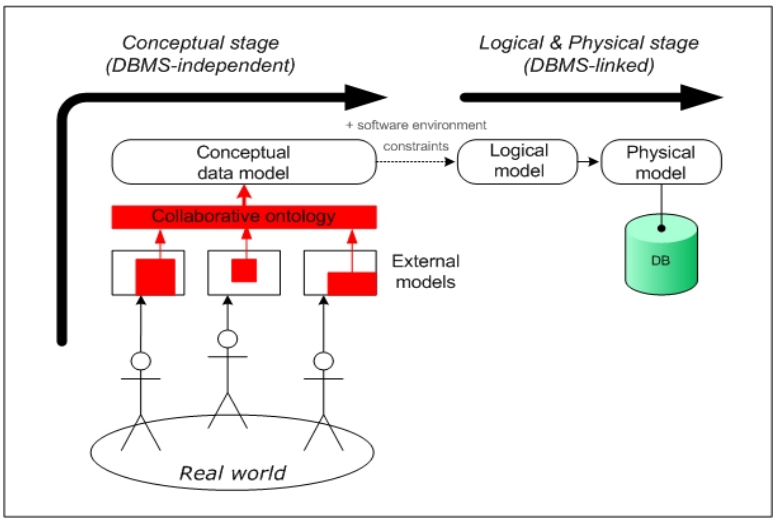


Fig. 1. Classical steps of SDB conception process. (from [6], p. 68, modified)

This stage forces the community of data producers, developers or even future users to (re-)think about the "basic geographic entities of their world" [7], regardless of any

database system. It is a natural step in the design process and this usually corresponds to some form conscious or unconscious of domain ontology design (Fig. 1).

Considering ontologies as a necessary step before the creation of tables and relationships is not new [8]. However, formalisation and storage of ontologies is not frequent in SDB and GIS design. It is probably due to underestimation of this highly conceptual stage (lack of knowledge), and the erroneous feeling of "loosing time" when concrete usable results are needed (feature catalogue, CDM, etc.). When urban ontologies have not been explicitly formalised, ontologies can be extracted *a posteriori*, knowing that the database is somehow based on an implicit conceptualisation (inverse order process).

Obviously such an extraction of ontologies from existing databases or GIS is especially relevant in the case of a reengineering of these information systems. As urban information is more and more available in digital format, reengineering is becoming a major concern for most institutions in charge of the maintenance of these data. Data reengineering may indeed be required by the present evolution of techniques (migration from one platform to another one, adoption of open-GIS format), of the requirements (new uses of the databases, increased performance requirements, web access, inter-operability) or the data itself (integration of new information sources, 3D extensions, use of automatic acquisition techniques). In any of these cases, ontology extraction from existing databases and GIS appears as a crucial step before addressing the technical issues of the reengineering process.

3 Spatial objects and relationships

Spatial objects have been formalised for a while (at least in 2D). In SDB and GIS, standardised spatial types are available (such as point, line, polygon, etc.). However, dealing with spatial information is much more than looking to spatial objects; it concerns also spatial relationships existing between them. In this matter, formalisation is far to be finished, even if standards have already been adopted for some type of spatial relationships. In GIScience, one distinguishes between qualitative and quantitative spatial relationships. The former ones do not refer to metrical concepts when the latter ones do. For example, saying that the city of Liège is {*disjoint of, not far from, east of*} the city of Brussels, is a qualitative statement, when saying that the city of Liège is *at 95 km* from the city of Brussels is a quantitative statement. Formalisation of such qualitative concepts is a key research in GIScience. Most of the work in the field has focussed on topological relationships. Such relationships are based on topological geometry and allow distinguishing relations such as "disjoint", "overlap", "included", … [9] [10]. These are far to be the only qualitative spatial relationships. However, we will restrict our discussion to them in this paper, as they are the only ones to be efficiently managed in SDB and GIS.

Beyond Egenhofer and Clementini operators, there are other ways to express topological relationships. For instance, the formalism CONGOO [6] considers two relations (Superimposition (S), Neighbourhood (N)) with three application levels:

total (t), partial (p), non existent (ne). For example, saying that Liège and Brussels are disjoint could be stated as: Liège S_{ne} N_{ne} Brussels.

This particular way to express topological relationships is equivalent to the ones adopted by the OGC, more information could be found in [6]. We will see that this geo-formalism has been selected for our case study and therefore it is worth mentioning some of its particularity. Beyond the expression of topological relationships, one of the main interests of the CONGOO is to propose the use of *topological matrices*. These matrices contain all the topological relationships that bind the object's sets together. There are two types of topological matrices; the *classical* and the *strong*. The classical matrix contains each topological relationship between every object with all the other objects. The strong matrix contains topological relationships which must exist between a given object and a given number of objects. Figure 2 illustrates the difference between both concepts.

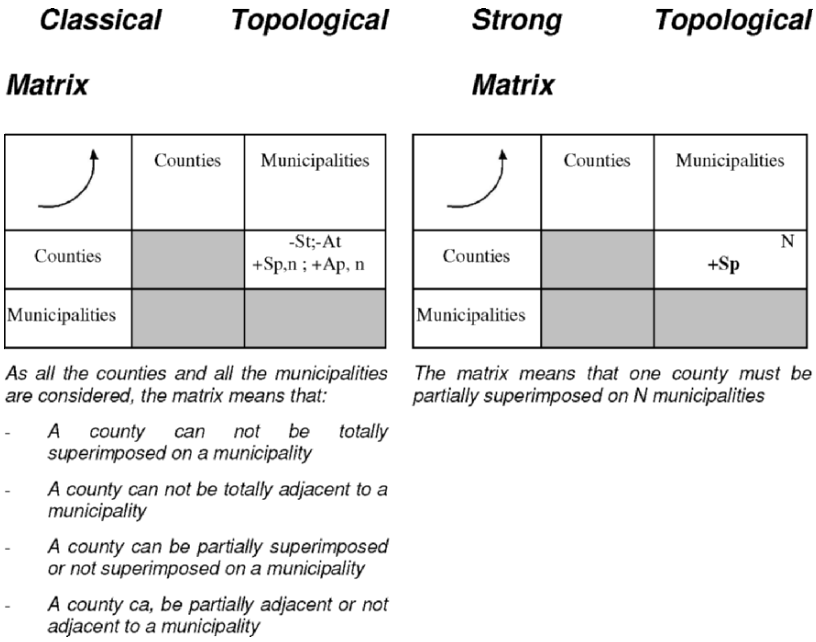
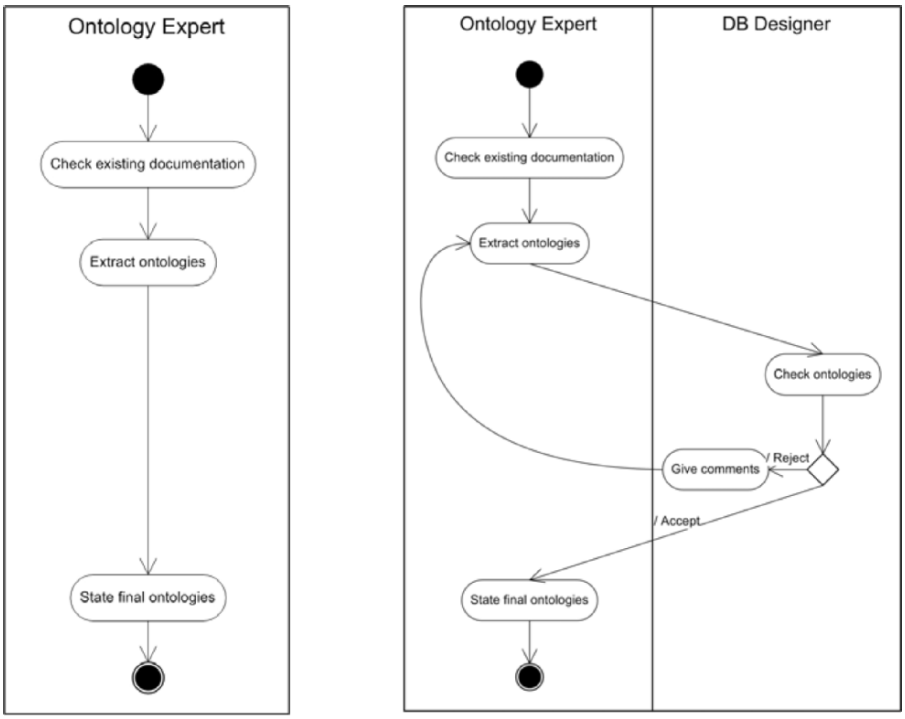


Fig. 2. Classical and Strong topological Matrices (from [11])

We do not claim in this paper to summarize topological relationships issues in one section. What is important to note is that such spatial relationships bring crucial information about objects spatial behaviours and consequently about spatial domain ontologies.

4 The bottom-up approach

The proposed bottom-up approach is rather simple and could be theoretically presented as follow (Fig. 3a).



a. Without DB designer b. With DB designer

Fig. 3. Proposed bottom-up approach (UML Activity diagram)

1. The first step is to analyse the existing database documentations and then extract a draft version of the ontologies. Local ontologies can be extracted from data catalogues or data dictionaries and semantic nets can be derived from CDMs (examples of extraction are presented in section 5). The derived ontology should be expressed in an ontology-language like KIF or OWL.
2. At this stage, two options are possible depending on DB designer collaboration.
   a. The relevance of extracted ontologies can be checked by comparing them to the related populated DB. Final ontologies can be then obtained and the extraction process ends.
   b. If it is possible, the next step is to submit the draft ontologies to the DB designer (then the bottom-up approach evolves to figure 3b). An important issue at this stage is to ensure that both "teams" use the same language, the same concepts. A definition is provided for each concept.

This definition includes a textual description as well as a formal expression of its relations with other concepts (IS A, part of and possible topological relations).

3. Remarks formulated by the "DB expert" team must be included in the ontologies extraction process and new ontologies have to be provided until final acceptation.

5 Case study: Brussels UrbIS 2

In Belgium, spatial databases are generally developed by the federal or regional administrations that manage and/or produce inventories of geographic data for the territory they are in charge of. Brussels UrbIS 2 © is the geographic information system of the Regional Government of Brussels.

At the end of the nineties, it became obvious that a complete reengineering of the databases was needed. A collaboration between the *Centre Informatique pour la Région Bruxelloise* (CIRB) and the Geomatics Unit of the University of Liege started in 1998 to provide the necessary support to achieve the reengineering process of part of the SDB (the ADM base containing 33 classes and 830000 instances mostly related to geographical administrative information), i.e. bringing the DB to its second operational version.

The objective of the first conventions was to create *a posteriori* a feature catalogue and conceptual data models. One of the first step was the (re)-definition of local ontologies of the original database [12]. This step has never been formalized for two reasons. Firstly the CIRB team was looking for quick and specific outputs, conceptual stage of the reengineering was not their priority. Secondly ontologies as part of the DB design process were not widely known in the GIS community at that time. Nevertheless, the bottom-up approach we have followed to extract these ontologies can be exposed.

5.1 Application of the bottom-up approach

The practical application of this approach has been rather difficult for several reasons.

First, the existing documentation was incomplete and non standardised. The only documentation available was some relational schemes, a data list (different from a catalogue structure) and data acquisition specifications (for photogrammetric and land surveying acquisitions). The geographical information contained in these schemes was rather poor. Only some hierarchic and thematic links have been deduced from them.

Second, the aim of the work was as we said the creation of DB feature catalogue and CDM, not explicit ontologies. Therefore, the submission process was not based on the validation of ontologies but on validation of these other outputs.

Third, the *database designers* (the CIRB team) failed at the beginning to validate the draft outputs. It was due to a misleading of conceptual perception of the geographical database. Therefore, we had to provide them the necessary tools and methods to formalize their knowledge. It implied to adopt a common language, and more

especially a common spatial language. For this purpose we have used first a "natural" language expressed within and Entity/Relationship (E/R) formalism, and later we adopted a more specialized *geo-formalism*. In the nineties, limitations of "traditional" formalisms for handling spatial information were highlighted and consequently several geo-formalisms were proposed: Modul–R [13], GeO - OM [14], MADS [15], CONGOO [6], Geo–UML [16], etc. to name but a few. Overall, these formalisms handle geographic representation of objects as well as spatial relationships. CONGOO has been selected because it was known by the experts in charge of the project.

The practical approach corresponds more to the next diagram (Fig. 4).

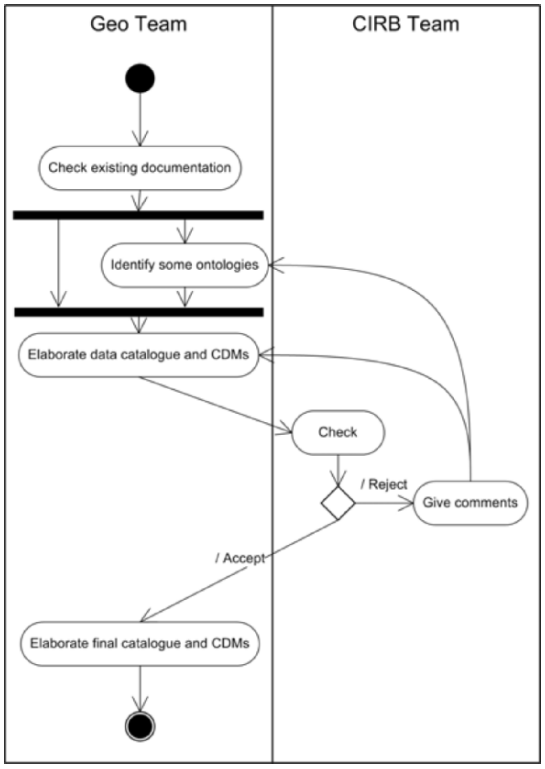


Fig. 4. Practical bottom-up approach (UML Activity diagram)

One of the most important aspects of the submission/acceptation process was the establishment of objects spatial properties: object representation and spatial relationships between objects.

In the reengineering project, topological matrices (*cf.* 3) had to be created with the CIRB team. This is the first step of any CONGOO conceptual model. The elaboration of this matrix is time consuming and a hard step of the process. Quite importantly, by

identifying spatial (topological) relationships between objects, this stage revealed object's definition inconsistencies. This point if further developed in the next section (*cf.* 5.2).

5.2 Ontologies extraction and objects definition

The aim of this experience was to define specific notions and then extract more generic concepts by semantic generalisation. The process started with proposing definitions for DB's basic objects, which should be very close to the ontologies that drove the DB's creation. However, one has to keep in mind that it was a reengineering project and therefore we could not ignore the complexity of existing DB's objects. The following example, presenting the evolution of the definition of the "house" object, illustrates the different levels of abstraction we had to consider.

The initial definition was clearly link with object's graphical construction and data sources (in this case the topographical survey).

Definition 1: The « house » is the building extract out of the topographical survey

We did not have the ability to change object's name, however, this "definition" was clearly not satisfactory. Our own understanding of the objects leads us to the following definition (whose validity was checked against other DB's documentation):

Definition 2: The « house » corresponds to footprint of a building (including its annexes)

This definition appeared to correspond to the designer ontologies. However, when considering spatial relationships between objects (from the topological matrix), the definition had to be adapted. The issue was highlighted when considering the topological relationship "superimposed to" (overlap). From objects definitions, it was expected that "house" could not be superimposed to object "street". However, CIRB team indicated that it was indeed possible because of the inclusion of objects such as bus stop, fountain, etc. into the object "house".

Definition 3: The « house » corresponds to building's footprint, including annexes and all other construction such as church, chapel, monument, school, fountain, greenhouse, bus stop, etc.

This definition is quite odd and not satisfactory conceptually. However, it corresponded to the reality of the DB and had been included in the feature catalogue. Of course, one of our DB's reengineering recommendations was to split this object "house" into several more semantically consistent objects.

In this example of the objects definition extraction, we can say that the definition 2 was indeed at a higher level of conceptualization. An ontological dictionary could have been produced at this stage, prior the feature catalogue. The project also raised linguistic issues due to the fact that the ontology had to be developed in both French and Dutch, which are the two official languages in Brussels.

5.3 Conceptual data models and semantic nets

As we have seen above, the extraction of ontologies during the reengineering process was a crucial step in the understanding of objects/concepts. As a logical step in the process, following data cataloguing, conceptual data models were built. One in Entity/Relationship formalism and the other with CONGOO (Fig. 5).

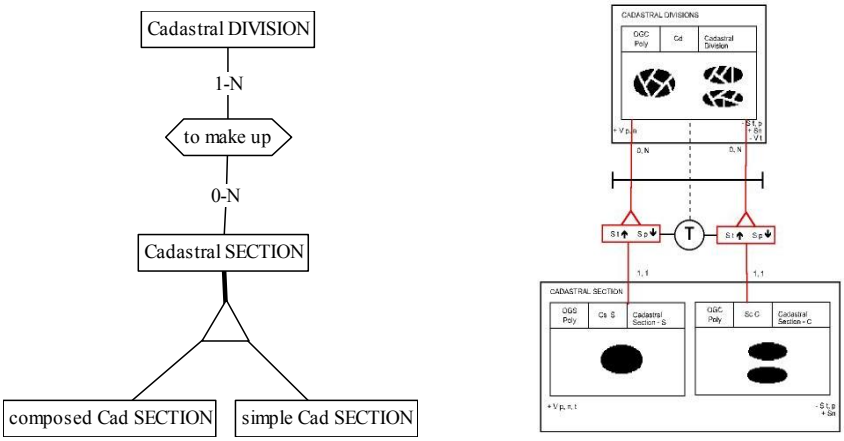


Fig. 5. Extracts of conceptual schemas: E/R and CONGOO

The temptation to (re-)interpret the E/R model as a semantic net is large. If it can become a really interesting and convenient synthesis and communication tool, such a schema is basically designed for a specific information system, describing the contents of a specific database, i.e. the specifications of one possible "world" [17], [12]. That means that we would have to operate an intermediate step to build a kind of semantic net (Fig. 6), based on the generic definitions. By this way, we would obtain a richer model (global-transposable-sharable) than the database conceptual schema, capturing the semantics of information in a formal way, and usable as a possible way for data integration [2]. This extraction process from E/R models can be envisaged (semi)automatically (selection of specific entities, relationships and attributes). It is not the case with CONGOO which is currently only a "graphical" formalism without CASE tools.

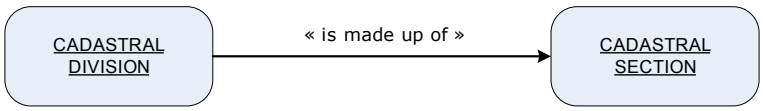


Fig. 6. Semantic net

Two "models" were thus proposed, the E/R model using complex types of relationships and specializing objects based on their geographical representations and the CONGOO model, much richer, including concepts such as classes and layers and representing all topological relationships between objects. It is worth noting that the CIRB team has continued to use and up-date the E/R model (the one closer to the semantic net) and has left behind the CONGOO one as they did not managed to maintain it.

6 Conclusion and future developments

In this article, we have discussed extraction of local SDB ontologies in the context of urban domain ontologies identification. We have tried to clarify the role of ontologies in SDB's design and reengineering. If the ontology level is necessary for DB's design (and interoperability) [18] [19], related ontologies are not always formalized. Therefore, local SDB ontologies are usually hidden in SDBs and associated documentations (feature catalogues and CDMs). In this case, it is possible to extract them from the documentation by applying a bottom-up approach. This process could be improved by a good collaboration with DB's original designer when the DB is poorly documented.

From our experience, extracting local ontologies (and associated objects definitions) implies a very good knowledge of spatial relationships between DB's objects. If extraction processes could be investigated further, it is clear that the major output of this reflective analysis is that local DB ontologies must be recorded during the DB design process. Another issue which should be tackled is the case of non documented DB or more generally non documented spatial numerical information; we believe that a comprehensive analysis of spatial relationships between instances could be the first stage of local ontologies extraction.

References

1. Gruber, T. *A translation approach to portable ontology specifications.* Knowledge Acquisition, 1993**5**(2): pp. 199-220.
2. Morocho, V., L. Pérez-Vidal, and F. Saltor. *Semantic integration on spatial databases.* in *Proceeding of VIII Jornadas de Ingenieria del Software y Bases de Datos.* 2003. Alicante. pp. 603-612.
3. Roussey, C., R. Laurini, C. Beaulieu, Y. Tardy and M. Zimmerman. *Le projet Towntology. Un retour d'expérience pour la construction d'une ontologie urbaine.* Revue Internationale de Géomatique, 2004. 14(2): pp. 217-237.
4. Van der Vet, P.E. *Bottom-up construction of ontologies.* IEEE Transactions on Knowledge and Data Eengineering, 1998. **10**(4): pp. 513-526.
5. Sowa, J. *Top-level ontological categories.* International Journal on Human-Computer Studies, 1995. **43**(5-6): pp. 669-685.
6. Pantazis, D. and J.-P. Donnay. *La conception de SIG. Méthode et formalisme.* 1996. Paris: Hermes, 339 p.

7. Fonseca, F., M.J. Egenhofer, P. Agouris and G. Câmara. *Using ontologies for integrated geographic information systems.* Transactions in GIS, 2002. **6**(3): pp. 231-257.
8. Gruber, T. *The role of common ontology in achieving sharable, reusable knowledge bases.* in *Proceeding of the International Conference on Principles of Knowledge Representation and Reasoning.* 1991. Cambridge. pp. 601-602.
9. Clementini, E., P. Di Felice and P. Van Oosterom, *A small set of formal topological relationships suitable for end-user interaction*. Advances in Spatial Databases LNCS 692, 1993. pp. 277-295.
10. Egenhofer, M. *A model for detailed binary topological relationships.* Geomatica, 1993. **47**(3-4): pp. 261-273.
11. Pasquasy, F., F. Laplanche, Jean-Christophe Sainte and J.-P. Donnay. *MECOSIG adapted to the design of distributed GIS.* in *On the move to Meaningful internet systems 2005*, R. Meersman et al. (Eds): OTM Workshops 2005, LNCS 3762, pp. 1117-1126.
12. Fonseca, F., C. Davis, and G. Câmara. *Bridging ontologies and conceptual schemas in geographic information integration.* GeoInformatica, 2003. **7**(4): pp. 355-378.
13. Caron, C., Y. Bédard and P. Gagnon. *MODUL-R, un formalisme individuel adapté pour les SIRS.*Revue Internationale de Géomatique, 1993. **7**(3), pp. 283-306.
14. Tryfona, N., D. Pfoser and T. Hatzilacaos. *Modelling behavior of geographic objects: an experience with the object modelling technique.* at *CASE*, 1997. Barcelona.
15. Parent, C., S. Spaccapietra, E. Zimányi, E. Donini, C. Plazanet, C. Vangenot, N. Rognon and P.-A. Crausaz. *MADS, modèle conceptuel spatio-temporel.* Revue Internationale de Géomatique, 1997. **7**(3-4): pp. 317-352.
16. Bédard, Y. *Visual modelling of spatial databases: towards spatial PVL and UML.* Geomatica, 1999. **53**(2): pp. 169-186.
17. Bishr, Y.A. and W. Kuhn. *Ontology-based modelling of geospatial information.* presented at *Third AGILE Conference on Geographic Information Science,* 2000. Helsinki.
18. Frank, A. *Spatial Ontology.* in *Spatial and Temporal Reasoning*, O. Stock, Editor. 1997. Dordrecht: Academic Publisher. pp. 135-153.
19. Smith, B. and D. Mark. *Ontology and geographic kinds*. in *Proceedings of the Tenth International Symposium on Spatial Data Handling*., T Poiker and N. Chrisman, Editors. 1998. Burnaby: Simon Fraser University. pp. 308-320.