Application of ER and CONGOO formalisms in a spatial database-reengineering project

Roland Billen[1], Dimos Pantazis[2] and Bernard Cornélis[3]

[1] Center for Geosciences, Department of Geography and Geomatics, University of Glasgow, Glasgow G12 8QQ, Scotland (UK), rbillen@geog.gla.ac.uk
[2] Department of Surveying, School of Technological Applications, TEI (Institution of Technological Education of Athens), Samarinas 3, Sepolia 10443 Athens, Greece, pantdonz@otenet.gr
[3] SPIRAL, Université de Liège, Belgium, cornber@yahoo.com

Abstract. This article discusses the application of a classic formalism (ER) and a geo-formalism (CONGOO) in the design stage of a spatial database reengineering. Through a real scale project, the advantages and disadvantages of both formalisms are presented as well as the different steps of the conceptual reengineering process. This study confirms the necessity of using a geo-formalism but does not reject the ER formalism as a first basic common language.

Forewords

This contribution stems from a project involving both the *Université de Liège* and the *Centre Informatique pour la Région Bruxelloise* (CIRB). The project consisted in an analytical study of a set of spatial and alphanumeric databases from *Brussels UrbIS 2*®© (Billen *et al.* 1998a, 1998b, 1998c, Sheeren and Billen 2000). As a result of the study, a data dictionary was developed, two conceptual models were elaborated, one with the Entity/Relationship (ER) formalism and one with the *CONception Géographique Orientée Objet* (CONGOO) formalism, as well as a set of data quality guidelines for the validation of the database's objects. This article illustrated by results deriving from the real scale project focuses on the conceptual stage of spatial database reengineering.

1. Introduction

In the nineties, limitations of "traditional" formalisms for handling spatial information were highlighted and consequently several geo-formalisms were proposed; Modul–R (Caron *et al.* 1993), GeO - OM (Tryfona *et al.* 1997), MADS (Parent *et al.* 1997), CONGOO (Pantazis and Donnay 1996, Pantazis 1995, 1997), Geo–UML (Bédard 1999), etc. to name but a few. The creation of proper conceptual models in early stages of spatial databases and GIS design is now widely accepted. Unfortunately for older spatial databases and GIS, conceptual models are usually missing. This lack of conceptual analysis becomes a major issue when such databases or GIS need to be reengineered. Shortly after starting the upgrade of the 10-year old *Brussels UrbIS* ®© project, it became obvious that this lack was hindering the development of the second version of the database. In the first stage of the reengineering of *Brussel UrbIS 2*®©, the proposed solution was to create conceptual data models of the existing database using ER and CONGOO formalisms. The ER model would give a first easy-to-understand view of the database, while the use of a geo-formalism was needed to ensure its complete description. CONGOO has been selected among others because it was an achieved object oriented geo-formalism (Bédard 1999) and because it was known by the experts in charge of the project.

The purpose of this paper is to study the usability of both formalism and geo-formalism in a reengineering conceptual stage when (re)-definition of ontologies of the original database is needed (Fonseca et al. 2003). In the first section, both formalisms are presented, briefly for the ER one and in more details for some aspects of CONGOO. The following sections expose the methodology followed in the different steps of the conceptual stage of the reengineering and present concisely a description of the two conceptual models. A discussion on the main points raised by this reengineering process conclude this contribution.

2. ER and CONGOO formalisms

2.1. ER Formalism – a brief recall

The conceptual ER model is a description of the real world human's perception expressed in terms of data structure as opposed to the physical organisation of files in the computer to ensure efficient storage and performance. This formalism is widely used in a range of methodologies (Merise methodology, James Martin Information Engineer, etc.). It is assumed that the reader is familiar with ER concepts such as entity, relationship and attribute. For this project, the selected formalism is the one proposed by Nanci and Espinasse (1996). This sub version is enriched by a generalisation/specialisation relationship and constraints between relationships (exclusion, completeness, totality, inclusion, or any combination of them) and by pictograms expressing the spatial form of the entity. Indeed, with these enrichments, the ER formalism used can be seen as a simplified version of the formalism Modul-R (Caron *et al.* 1992).

2.2. CONGOO's concepts

Geo-formalisms take into account the specificity of spaces in the human conceptions of the real world (Jones et al. 1996, Freundschuh and Egenhofer 1997, Renolen 2000). Since the use of CONGOO formalism is far more marginal, it seems essential to give a more detailed explanation on its basic concepts and to provide some useful extended concepts which have been intensively used in this real scale project.

Basic concepts. CONGOO defines and distinguishes *classes*, to be understood as 'set', 'group of objects', and *objects*. The description of an object is provided by properties called 'attributes', by behavioural characteristics named 'treatments' and/or 'methods' (in case of complex treatments), and by structural characteristics concerning the topological and other types of 'relationships' between the objects, classes, layers and sub-layers (see further for definition of the terms 'layer', 'sub-layer').

The following object categories are distinguished:

- The *geo-graphic objects* (GO). The geo-graphic objects are the objects having a graphical representation spatially/geographically referenced. They are divided in three sub-categories covering the whole range of geo-graphic objects at the conceptual level:
- – Simple geo-graphic objects (SGO), gathering three sub-sub-categories:
    1. type point (SGO, type P)
    2. type line (SGO, type L)
    3. type surface (SGO, type S)
- – Composed geo-graphic objects (CGO). This kind of object stems from the union or the division of simple geo-graphic objects of the same type, belonging to the same or different classes.
- – Complex geo-graphic objects (CXGO). This category is made of geo-graphic objects that stem from the union of different types (points + lines, lines + surfaces, etc.) of geo-graphic objects (simple, composed or complex or a combination of them).
- The *non geo-graphic objects* (NGO) concern all the other kinds of objects (alphanumerical, aerial photos or satellite images without geographic references, video data, etc.).
- *Attributes.* Every object may have two kinds of attributes: the 'graphic' attributes (which relates to its type and its graphic representation) and the logical attributes (which concern its identity description). The non-geographic object and the classes only have logical attributes.

- The *classes* are sets of objects that have common characteristics expressed by their logical attributes and/or their common name (semantic identification). The graphic characteristics and the implementation mode (point, line, surface, combination, etc.) can be different for objects belonging to the same class.
- The *layers* are sets of classes of objects that have in common different kinds of relationships.
- The *sub layers* are sets of classes of objects of the same layer that have in common different kinds of relationships.
- The *relationships* are divided into structural, topological and logical relationships.
  - *Structural relationships* come under the 'Generalisation-Specialisation' (Gen-Spec) and the 'Whole-Part' types.
  - *Topological relationships*: Every geo-graphic object and every class of geo-graphic objects is described in space compared to other objects and/or other classes of objects, sub-layers, layers through two topological relationships: the 'adjacency' [A] and the 'superimposition' [S] relationships. Moreover, the 'obligatory' (must be [+] and/or forbidden[1] [-]) and the 'allowed' (may be; permitted) topological relationships are distinguished as well as three application levels: total [t], partial [p], non existent [ne]. All references to topological relationships in this article should be understood as the topological relationships according to CONGOO. The topological relationships must be maintained in the conceptual data model but also in a matrix, called the topological matrix, which contains all the topological relationships that bind the object's sets together. The creation of this matrix is a very important step in the design process.
  - *Logical relationships* concern all the relationships that are neither topological nor structural, e.g. the fact that a factory (geographic object represented by a point) that rejects its waste water into a river (geographic object represented by a line) is an information that will relate the two geographic objects (river and factory) but is not a topological (the factory-point and the river-line have no common topological relationship) nor a structural relationship (as defined in this paper).

Extended concepts. One of the main reasons to make use of a geo-formalism such as CONGOO is to express without ambiguity the topological relationships between geo-graphic objects. Knowing that a conceptual model aims at being exhaustive and readable, one can easily understand that an exhaustive representation of all the topological relations between objects would make the model inefficient. CONGOO proposes some solutions to meet these contradictory objectives by reducing the number of drawn relationships.

In CONGOO the topological relationships can be expressed in different ways in the model. They can be drawn between objects, between classes, between layers, or appear in the classes and in the layer. Figure 1 presents the different ways to express topological relationships.

[1] We underlay the possibility to describe forbidden relationships, which are usually important for the spatial database design.

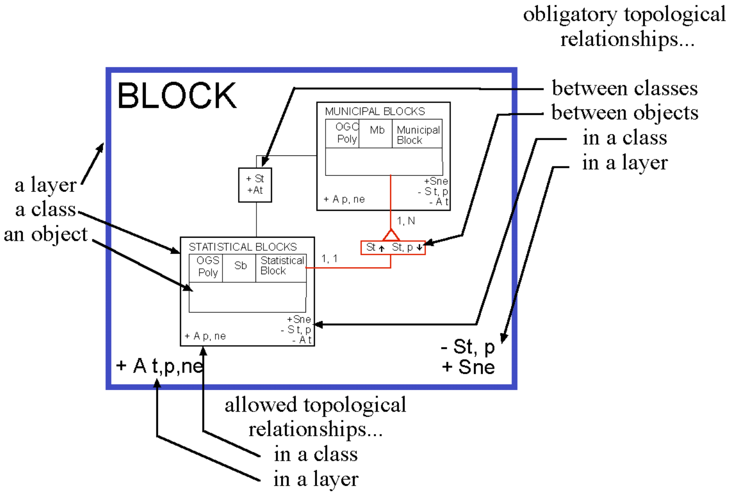

Fig. 1. Different ways to express topological relationships in CONGOO

Topological relations drawn between objects. In CONGOO, topological relationships between objects can be drawn between these objects (such as relationships between entities in ER). This kind of relationships always takes precedence over all others kinds. Sometimes, it is necessary to use cardinalities with the topological relationship between two objects. It means that the relationship is valid only for the number of instance specified with the cardinalities. It is often the case when there is a whole-part relationship between two geo-graphics. In this case, the topological relationship is drawn with the structural relationship (figure 1).

Topological relations drawn inside classes and layers. A topological relationship present in the corner of a class (or a layer) will affect all the objects of the class (or all the objects and classes of the layer) and will be mandatory if drawn in the bottom right corner or allowed if drawn in the bottom left corner. Such a representation allows reducing the number of drawn relationships but of course requires the existence (or the creation) of classes or layers (figure 1). It appears that, if the relations between objects allow this and if the layer is well designed, the use of the CONGOO formalism reduces the number of drawn relations but this reduction is only graphical; the relations themselves are still present. An example of this aspect of CONGOO is presented here. The three 'not superimposition' ER relationships between the 'Urban block', the 'Street' and the 'Square' objects (figure 2) are replaced by one CONGOO topological relationship within the new layer URBAN COVER (figure 3).

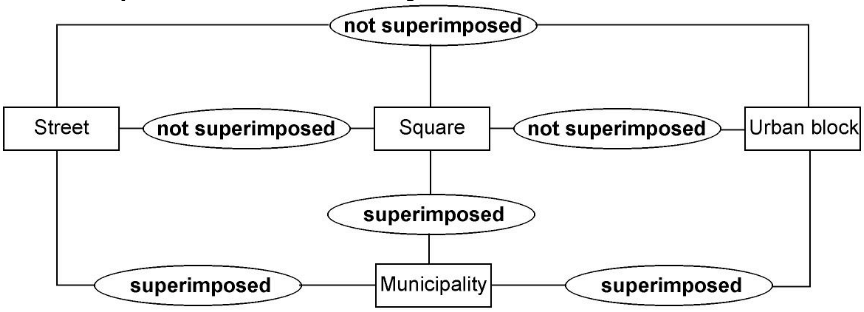

Fig. 2. Some relationships between *Brussel UrbIS 2®©* entities (ER model)

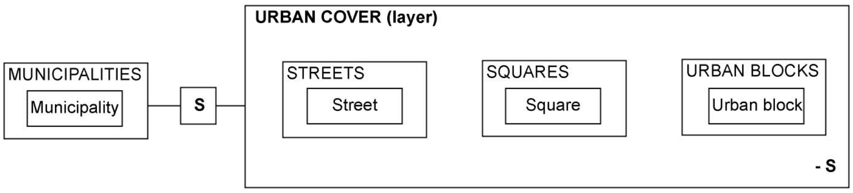

Fig. 3. Corresponding model with objects, classes, layer and topological relationships (simplified CONGOO model)

Topological relations drawn between the classes, between the layers and between classes and layers. Another possibility to reduce the number of drawn relations on the scheme is to consider topological relationship between layers, between classes and between classes and layers. This kind of relation is frequently used in CONGOO and can be interpreted like this: 'each instance of each object of one class (or layer) is in relation with any instance of any object of the other class (or layer)' (see Pantazis and Donnay (1996) and Pantazis (1997) for a complete interpretation). The concept is illustrated in the figures 2 and 3. The Municipality has three relationships with the other entities in the ER model (figure 2), which are replaced by one topological relationship between the class MUNICIPALITIES and the layer URBAN COVER (figure 3).

The combination of these three types of relationships representation allows, when the structure of the model is optimal, to reduce drastically the number of drawn relationships. In other words, the better the model is structured in terms of set of objects, classes and layers, the more readable it will be.

The expected number of topological relationships. The CONGOO formalism admits two topological relationships: 'adjacency' [A] and 'superimposition' [S]. With the distinction between the three application levels (total, partial, non existent), six relationships can be found: A t, p, ne and S t, p, ne. Other formalisms usually consider one relationship to express the topological situation between two objects. Therefore, for the sake of a sound comparison, we will consider that only one topological relationship represent the topological configuration between two objects and therefore the total number of relationships that is expressed in a complete CONGOO model including *n* objects is equal to n^2.

3. Conceptual stage of the Reengineering – ER model

3.1. Goal of the E/R conceptual model.

The purpose to design an E/R model was to provide a first description of the entities, to list all their attributes, to show some of the important relationships between entities and to give a first conceptual representation of the database.

3.2. Design process

Before reengineering, the only documentation available was some relational schemes (linked tables) and a description of the database itself. The relational schemes (linked tables) were related to the implemented relational model. The geographical information contained in these schemes was very poor. Only some hierarchic and thematic links have been deduced from them. The most interesting source of information was the practical experience and knowledge of the *database designers* (the CIRB team). Unfortunately, their conceptual perception of the geographical database was incomplete or at least not well structured. The method used to elaborate the conceptual model was iterative and a lot of contacts were needed in order to define the entities and the relationships. During the first analysis, the *expert team* (Univeristé de Liège) aimed at understanding how the database had been designed. A first model was made and proposed to the CIRB team. Step by step, the CIRB team refined their perception of the entities and relationships. After a few iterated models, they could give more and more detailed descriptions of their conceptual view of the database. The final model, derived from the collaboration of both teams, is a clear conceptual view of the database at a specific time. Its structure is still linked to hierarchic and thematic criterions. This conceptual stage of reengineering allowed the expert team to clearly understand the database.

3.3. Description.

The elaborated E/R model contains 35 entities, 12 constraints, and 43 relationships of which 13 are whole-part relationships, 8 generalisation/specialisation relationships and the 22 remaining ones express other relationships (an overview is presented in figure 4).

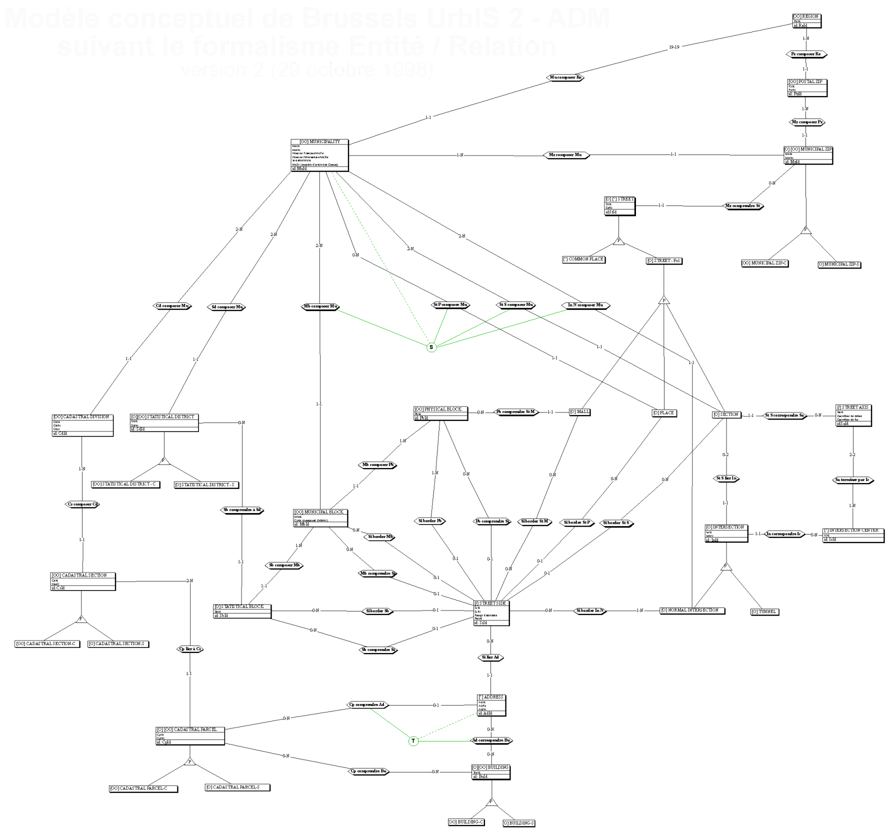

Fig. 4. An overview of the ER model of *Brussel UrbIS 2*®©

Even if the ER model has achieved its aim, it can not obviously be considered as the final output of the conceptual stage. It presents too many lacks in handling geo-information especially in the management of topological relationships. Even if it is possible to express topological relationships as ER relationships (using “adequate” expression which is however relatively imprecise), the number of relationships that have to be established between all the entities is too large and the model quickly becomes unreadable (see for example Detienne, 1996).

4. CONGOO conceptual model

4.1. Goal of the CONGOO conceptual model.

The objective of this model was to obtain a complete model where all the relations and all the objects of the database could be found. The *Brussels UrbIS 2*®© CONGOO model should be more than a single picture of the database at a given time. It should be a conceptual dynamic component of the database. All the changes that could affect the database must be integrated into the CONGOO model to avoid inconsistency in the physical database.

4.2. Model design process.

On the basis of the ER model, it was possible to dress the list of all the geo-graphics of *Brussels UrbIS 2*®©. At that moment, the creation of a database dictionary (Pantazis *et al.* 2002) and of the CONGOO model began. The topological matrix was created with the CIRB team. This matrix contains all the topological relationships that bind the object's sets together. 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. Each topological relationship between every object with all the other objects must be indicated. On the base of this matrix, the objects can be grouped following a topological criterion (instead of semantic criterion). A first model exclusively based on topological criteria was achieved in two steps. First, an automatic grouping program based on a similarity classification algorithm is applied. This program allows to group object together according to their topological character. At the end of this step, the final grouping can be processed. Indeed, the algorithm groups preferentially the objects of same implantation (point, line, and surface). Thus, it is necessary to refine the first step result. This two-step process allows getting the layer structure of the model. The outcome brought a new vision of the database. When the first model was created, it was submitted to the database designers. Their comments and remarks allowed to refine the model and then to achieve a CONGOO model of the database at a given time.

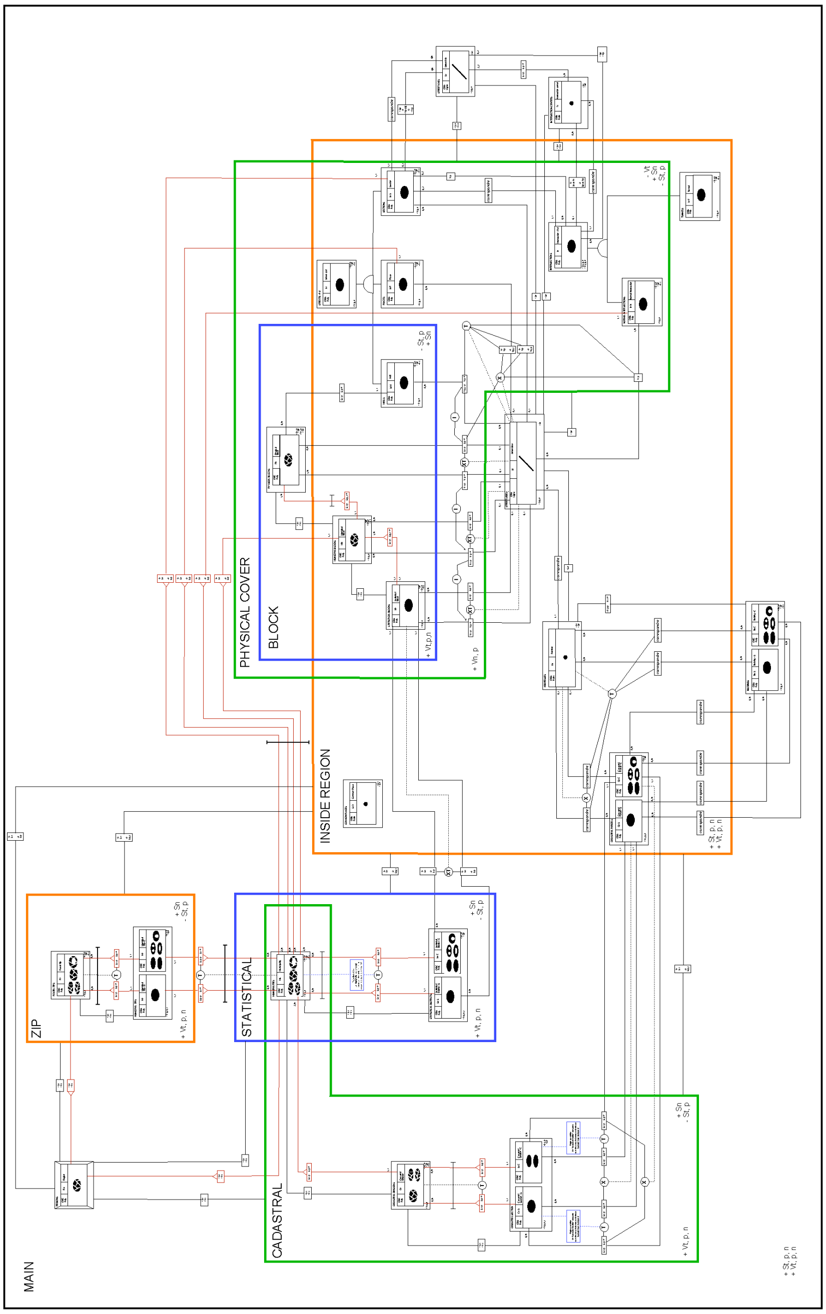

Fig. 5. An overview of the CONGOO model of *Brussel UrbIS 2®©*

4.3. Description of the CONGOO model.

The CONGOO model contains 29 objects, 24 classes, 1 sub-sub-layer, 5 sub-layers, 1 layer and 18 constraints. It also contains 872 relationships of whose 19 are structural relationships (17 whole-part and 2 generalisation/specialisation), 12 are logical relationships and 841 are topological relationships (an overview is presented in figure 5).

As an object-oriented geo-formalism, CONGOO proposes a wide range of tools and concepts for handling geo-information. In this real scale study, the following characteristics were particularly important:

- The different kinds of relationships (structural, logical, and topological) are each drawn in a different way, so it is easy to distinguish them.
- All the topological relationships and constraints between all the geographic objects are represented with only two concepts: adjacency and superimposition in a very intuitive manner. This provides the possibility for a total representation of an 'integrated topology', meaning that all intersections (= total or partial superimpositions), containment (total superimposition, total adjacency) and adjacencies are represented. A huge amount of topological relationships can be managed easily.
- The concepts of object, class and layer allow representing a clear and complete structure of the model.
- The integration of class and layer concepts in the management of topological relationships provides a powerful tool for the improvement of the readability of the model. The model contains 872 relationships[2] and is still very clear.

Nevertheless, CONGOO has some disadvantages. Its use is marginal, and for most of the designers, is not well known. Then, they often hesitate to invest in a new, and more complex formalism knowing that it requires a long training period. Another current limitation is the non-existence of CASE-tool. Hence the automatic translation into a logical data model is not possible.

5. Conclusions

The major issue in a reengineering process is that, obviously, the existing situation has to be taken into consideration. In other words, it is not a *from-scratch* process, and this can be quite complex if documentation (data dictionary, conceptual models) is poor or missing. A clear identification and definition of the objects and their relationships are needed and usually result from long discussions with the database designers, owners and users. These actors are not at all familiar with geo-formalisms and spatial concepts, and therefore it is easier to propose, in a first approach, an easy common language which will be understood by all. This is why it was decided to use, prior to a geo-formalism, the ER formalism, which despite its limitation for handling geo-information, has achieved this objective. It also had a strong didactic effect on the database designers by persuading them of the usefulness of the conceptual stage and more specifically of the usefulness of a geo-formalism to integrate all the complexity of geo-information. In the future, one can think that such a stage should not be anymore necessary if relevant professionals become more familiar with spatial concepts and geo-formalisms. The CONGOO model has fulfilled its role all along the design process. Some topological contradictions were found and the database designers directly addressed solutions. Most of the time, these solutions consisted in erasing or creating new objects or refining the existing definitions. Thus, the creation

[2] When considering the 6 CONGOO topological relationships, the number of relationships effectively present within the model is 4614.

of a clear and exhaustive conceptual data model gave important directives to the database designers. The result of this study confirms that the use of a specific geographical formalism is required to design geographical database conceptual models, without erasing the possibility of starting with standard formalism like ER for basic communication purposes.

Acknowledgments

The authors would like to thank the kind support of the CIRB staff, who closely collaborated with us in discussing various aspects of the data models and dictionary. Our recognition goes in particular to Mr Vanderborght, Mr Van Acker, Mrs Himpe, Mrs Roland and Mr Delande.

References

Bédard, Y., 1999, Visual modelling of spatial databases: towards spatial PVL and UML. Geomatica, 53(2), pp. 169-186.

Billen, R., Cornélis, B., Muller, F., Pantazis, D., Thiam, S. and Donnay, J.P., 1998a, Élaboration du modèle conceptuel de données de la base UrbIS2 ADM - version 1, SURFACES, Université de Liège, Scientific and Technical Report N°13.

Billen, R., Cornélis, B., Muller, F., Pantazis, D., Thiam, S., and Donnay, J.P., 1998b, Ébauche du dictionnaire de données de la base UrbIS2 ADM - version 0.1α, SURFACES, Université de Liège, Scientific and Technical Report N°13 bis.

Billen, R., Pantazis, D. and Donnay, J.P., 1998c, Validation des modèles conceptuels de données de la base Brussels UrbIS 2 ®©, SURFACES, Université de Liège, Scientific and Technical Report N°16.

Caron, C., Bédard, Y., and Gagnon, P., 1993, MODUL-R, Un formalisme individuel adapté pour les SIRS. Revue internationale de géomatique, 7(3), pp. 283-306.

Detienne, V., 1996, Analyse conceptuelle de documents cartographiques. Institut d'informatique, Facultés Universitaires Notre-Dame de la Paix, Namur, 165 p.

Fonseca, F.T., Davis Jr, C.A. and Câmara, G., 2003, Bridging ontologies and conceptual schemas in geographic information integration, GeoInformatica, 7 (4): 355-378.

Freundschuh, S.M. and Egenhofer, M.J., 1997, Human conceptions of spaces: Implications for GIS, Transactions in GIS, 2 (4): 361-375.

Jones, C.B., Kidner, D.B., Luo, L.Q., Bundy, G.L. and Ware, J.M., 1996, Database design for a multiscale spatial information system, International Journal of Geographic Information Systems, 10 (8): 901-920.

Nanci, D., and Espinasse, B., 1996, Ingénieurie des systèmes d'information: MERISE deuxième génération 3rd edition (Paris: Sybex).

Pantazis, D.N., 1995, CONGOO formalism basic concepts and three conversion examples, in Proceedings of the First Joint European Conference and Exhibition on Geographical Information, Eurogi, Vol. 1, The Hague, the Netherlands, pp. 223-226.

Pantazis, D., 1997, CON.G.O.O. : A conceptual formalism for geographic database design. In Geographic Information Research, Bridging the Atlantic (London: Taylor & Francis), pp. 348-367.

Pantazis, D., Cornélis, B., Billen, R., and Sheeren, D., 2002, Establishement of a geographic data dictionary: a case study of UrbIS 2©, the Brussels regional government GIS. Computers, Environment and Urban Systems, 26, pp. 3-17.

Pantazis, D., and Donnay, J.-P., 1996, Conception des SIG Méthode et formalisme (Paris: , Éditions Hermès).

Parent, C., Spaccapietra, S., Zimányi, E., Donini, P., Plazanet, C., Vangenot, C., Rognon, N., and Crausaz, P-A., 1997, MADS, modèle conceptuel spatio-temporel. Revue Internationale de Géomatique, 7(3-4), pp. 317-352.

Renolen, A., 2000, Modelling the real world: Conceptual modelling in spatiotemporal information system design, Transactions in GIS, 4 (1): 23-42.

Sheeren, D., and Billen, R., 2000, Méthodologie de validation de la qualité des données de la base de données UrbIS 2 - ADM, SURFACES, Université de Liège, Scientific and Technical Report N°21.

Tryfona N., Pfoser D., Hatzilacaos T., 1997, Modeling behavior of geographic objects: an experience with the object modeling technique, CASE'97, Barcelona.