

UPDATING DATA IN GIS : TOWARDS A MORE GENERIC APPROACH

Par Hakima KADRI-DAHMANI
COGIT¹ – LIPN²

Résumé

Assurer l'actualisation des données dans un système d'information géographique est une tâche des plus ardues. En effet, le caractère temporel des phénomènes naturels (une rivière qui sèche) ou non (changement du trajet d'une route) qu'elles représentent nécessite une révision régulière si ce n'est continue de l'ensemble de ces données. Le coût exorbitant de cette révision est le premier problème que rencontrent les organismes chargés de cette opération. En général, ces organismes sont eux-mêmes les producteurs et les fournisseurs de ces données et doivent cette actualisation à leurs clients ou tout simplement aux utilisateurs de leurs données. Leurs bases de données étant constituées sans prise en charge des problèmes [Badard : 97] que peut engendrer cette mise à jour, ils sont obligés de trouver des solutions pour répondre à certaines questions qui se posent: comment minimiser le coût de la révision [Bréart :00], comment est-il possible de caractériser les mises à jour effectuées [Raynal :96], quelle est la meilleure façon de livrer ces mises à jours réalisées ou encore comment est-il possible d'intégrer et de propager ces mises à jours dans d'autres jeux de données [Bedard :97][Harrie :99] [Uitermark :98] entre autre dans les jeux de données en possession des utilisateurs. C'est alors qu'il y a eu plusieurs travaux essayant de répondre à l'ensemble de ces questions, certains se basant sur les résultats ou les concepts des bases de données classiques, d'autres sur leurs expériences sur le terrain. Pour ce dernier cas, nous référons particulièrement les travaux effectués à l'IGN (Institut Géographique National) de France [Badard :99] [Lemarie :99].

Nous observons un point commun entre l'ensemble de ces travaux ; ils ne sont pas génériques et même s'ils répondent bien à l'une ou à l'autre de ces questions, ils laissent les autres sans réponses ce qui n'est pas sans engendrer d'autres types de problèmes. Par exemple, si l'on se focalise sur le problème de la représentation de l'information d'évolution dans le but d'une détection plus facile de cette information et que l'on laisse la tâche de propagation à l'opérateur [Dell'Erba :97], nous tombons dans le problème de l'imprécision de l'homme avec toutes ses conséquences.

Nous pensons que ce manque de généralité dans les solutions est dû principalement au fait que la conception des systèmes d'information géographique néglige la prise en compte de l'évolution de cette information.

Principalement, c'est ce que nous essaierons d'explicitier dans cet article. Nous présentons en premier une revue des diverses causes d'une mise à jour dans les bases de données géographiques ainsi que les problèmes qu'elle peut engendrer. Par la suite et sans qu'il soit exhaustif, nous présentons un état de l'art des travaux qui se sont intéressés à l'évolution des données géoréférencées et leur mise à jour dans les bases de données géographiques ; quel sont les points pris en charge par chacun d'entre eux, les points négligés et les conséquences qui s'en suivent. Nous finissons l'article en soulignant l'avantage de prendre en charge l'évolution de l'information géographique dès la conception du système.

Abstract

This paper deals with the study of geo-information evolution; the diversity and the complexity of the problems that arise are addressed. We present a brief state of the art of works related to the updating process in geographic information systems. We discuss what are their respective contributions and specific weak points. Finally, we introduce the generic solution we perceive to the problem of geographic database updating. We also justify the advantage of this solution which relies on the fact that the evolution of the geographic information is taken into account concurrently to the design of the system and not afterwards. Our challenge is to provide a computational model that makes the maintenance of such systems easier.

¹IGN, 2-4 avenue Pasteur
94 165 St Mandé Cedex , France
Fax : +33 1 43 98 81 71, hakima.kadri-dahmani@ign.fr

²Laboratoire d'Informatique Paris Nord, Avenue Jean-Baptiste Clément
F-93430 Villetaneuse, France
Fax : +33 1 48 26 07 12, hkd@lipn.univ-paris13.fr

Introduction

Updating is an essential step in GIS life cycle. If data are not regularly updated the results and decisions deduced from spatial analysis are unreliable. However, realizing this step is a very difficult task. Due to the temporal property of the natural (a dry river) or non natural (a road change) phenomena they represent, data require regular if not continuous updating. This compels data suppliers (such as IGN the French National Mapping Agency) to regularly update their databases and to update also all the products derived from these bases (e.g. cartographic dedicated products, other databases with different scales or possibly previous versions of user databases).

While these databases were usually conceived without taking this temporal property of spatial information into account, suppliers today must propose solutions to a variety of issues: How to minimize the cost of the updating operation? How to formalize the different updating operations to reduce interactive work [Egenhofer and al 92] [Raynal 96]? How to integrate the updating in the derived products [Uitermark and al 98] [Harrie et al 99]? Or even what is the best delivery mode for the updating information [Bedard and al 97]? and how to preserve the consistency of databases when the data modifications have to be integrated and propagated in the whole database? In recent years there has been a body of research tackling these problems. Some of them propose solutions which are inspired by classical (not geographical) database concepts. The others rely on practical experience such as research at the French NMA [Badard 00].

It may be underlined that the common weak point of these different works is their lack of genericity. Indeed, although they precisely address one or several of these questions they left the others unanswered.

The aim of this paper is on one hand to show that the main cause for these passed researches weak point comes from the poor ability of habitual GIS conception methods to handle evolution of geographic information, on the other hand to give first elements of what we consider to be a generic solution to the updating problem in GIS life cycle.

Updating data in GIS

Changes in the geographical reality leads in a new state of reality. From the mapping of these two states (old, new) of reality we get data for updates. These updates must be integrated into geo-databases and also in all products derived from these databases. Before they are integrated into the derived products, these updates go through several steps which we classify in three main steps: update integration in the source database, update integration in the derived products (diffusion) and the update exchange. The following figure illustrates this classification. Several researches have been devoted to overcoming problems met in each of these steps. The purpose of this section is not to draw up a complete review of different solutions defined in the literature, but rather, to present for each step some solutions and to illustrate their lack of genericity.

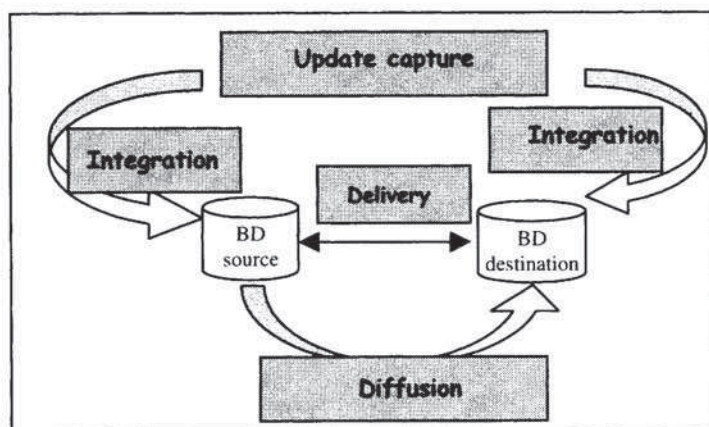


Figure 1 Main steps of the updating flow

Update capture

Update capture is the first step in the chain. This update can have several causes [Badard 00]:

- Real world evolutions : all spatiotemporal evolutions which concern the entities of the real world modeled in the geographic databases.
- Errors corrections : they correspond to correc-

tions of the ancient (former) version of the database to reflect in best the reality ground and this even if there were no real evolutions. It can concern the correction of an error arisen during the representation of this reality.

- Specification changes : the designer of the system may find necessary to change the conceptual model of his system to make it richer and/or of better quality. This requires a change of specifications, which are specifications of capture or contents.

The frequency of the collection, and consequently the

frequency of revision of bases, is the first problem at the producers of the geographic information [Bréard 00]. Some apply processes of revision by lot which consists in revising each part of the base at regular intervals. But the definition of the adequate interval for the revision of the adequate part is itself a subject of discussion. Certain themes of the base (for example the road theme) evolve much more quickly than the other themes (for example administrative limits). The different update cycle for the various themes, for example six months for the road theme and two years for the administrative limits theme, can engender incoherence in the base. If a road represents the administrative limit of a municipality and the course of the road is modified but not the administrative limits of this municipality then the base becomes inconsistent when share of geometry is expected. One could think that a continuous and homogeneous revision of all the themes of the base is the solution, which however may turn out to be very expensive and especially useless for certain themes of the base.

Update Integration

The update integration in a geographic database deals with the modification of all the entities directly or indirectly concerned by the evolution (e.g., splitting of a wood in several parts, which is crossed by a new road through). This task relies on retrieval of different relationships between objects and on the semantics of entities stored in the database. These relations are often implicit (i.e. not explicitly stored in the database) and have to be retrieval on the fly. Several works propose various methods to establish these relations such as the method of matching proposed in [Badard 00] and [Bonani 98] or the mechanisms of history in [Spéry 99]. In [Uitermark and al 98] the matching process depends on the abstraction rules.

Several kinds of conflicts may emerge when an update is integrated in a database. These conflicts provoke the inconsistency of the base. In most of the approach, to remove these conflicts need an operator intervention. Figure B.2 shows that the integration of the road creation provokes some conflicts.

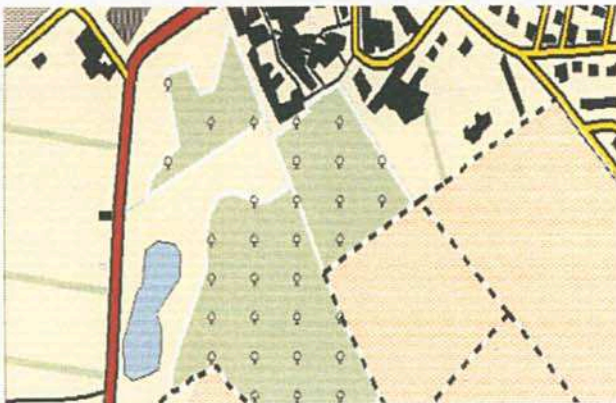


Figure 2.a Initial Database.

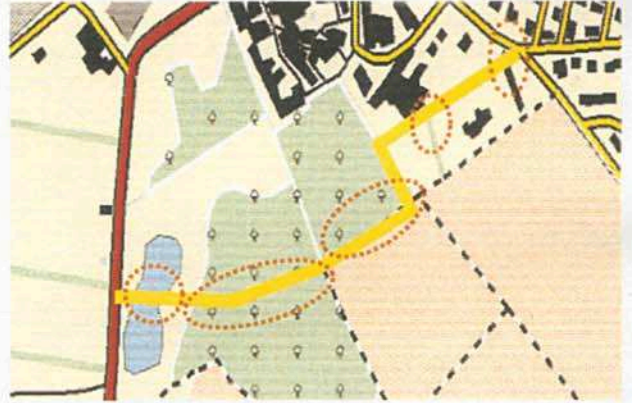


Figure 2.b Modified Database.

Figure 2 Road integration in the database causes conflicts

Update Delivery

Mostly, the update delivery is made in the form of the whole updated database. To integrate these modifications into a destination database, we have the choice between two alternatives: replace completely the destination database by the delivered one, or extract just the modified objects from the delivered database and integrate them in the destination one. The advantage of the first method lies in its simplicity of execution: it is enough to transfer all the current base of the producer. But, to extract the modifications and to propagate them in the own base constitutes a long and boring labour. The delivery by differentials implies that only the modified objects, between two states of the database are delivered. This approach is very advantageous because it allows to pass on only the data which should be updated, without affecting the other data. Furthermore it mitigates the problem of the quantity of information to be passed on because this quantity becomes

lesser and in that case the transfer is easier to make. The inconvenience of this method of delivery is that it is heavy and difficult to set up. Another type of delivery is what one calls logs [Badard 00]. In this mode, one delivers no object but only the description of the evolutions which objects (tracked down by their reference) underwent. This shape of delivery can concern objects having evolved but also whole base. The major inconvenient of this solution is that supposes the reference existence, what is not always acquired.

Diffusion of updates

To diffuse an update consists in using an update of an initial database (or a database source), from which updates to be made in another databases are deduced. For example, the diffusion of the updates of the BDTopo® (1:25000 database with one meter resolution produced by the IGN) to the TOP25® (a database derived from

BDTopo®). It can be made between a database and a derived product or between two independently surveyed databases.

Several kinds of conflicts exist between different databases, for example different models, schemas, classes and data structures. With spatial information there are extra conflicts, for example different geometries (polygons Vs polylines), different segmentations (road Vs road segments) and different aggregations (houses Vs building blocks).

One of the interesting studies which focus on update diffusion is reported in [Kilpelainen 95]. Kilpelainen's work addresses the issue of maintaining multiple representation in geo-database. It defines an approach that uses operations of generalization to update a geographic database from an other geographic database at a higher scale. This approach is called incremental generalization. Incremental generalization requires that the problem can be divided into "modules". Kilpelainen (1995) defines a module as a geodata entity that can be processed independently from its surroundings. Generalization is then performed only for the modules influenced by the updates. This updating method supposes the knowledge and the automation of the generalization process then it can be performed only in this specific case of multi-scale databases. Another problem is how to find the border line for a module: identification of modules is a crucial point but it is far from being easy to implement.

An other interesting approach is defined in [Uitermark 98]. It addresses the propagation of updates between two independently surveyed databases. The idea is to establish a knowledge base which contains a set of "Abstraction rules" and to "synchronize" both databases according to this knowledge base. Then to establish the relations between the objects of the two bases to find corresponding objects. Finally, using these correspondences, to integrate the updates in the second base.

This approach emphasises the relevance of defining a strategy to perform a consistent and complete propagation of updates between databases with different scales. However full scale update propagation depends on the knowledge of the abstraction rules of the data. Bases providers should express their abstraction rules in a shared vocabulary.

In the two previous studies, the updates are assumed to be clearly identified, which is far from being a general case. In [Badard 00] a mechanism dedicated to the update diffusion is proposed, in which updates are not identified: a process for automatic retrieval of updates between two databases is included. Figure3 illustrates the general structure of this updating mechanism. The approach defined in [Badard 00] seems more generic but has been implemented and validated only on databases produced by the IGN and more precisely, just between a database and its derived product (even if a part on interactive work is involved in the derivation process).

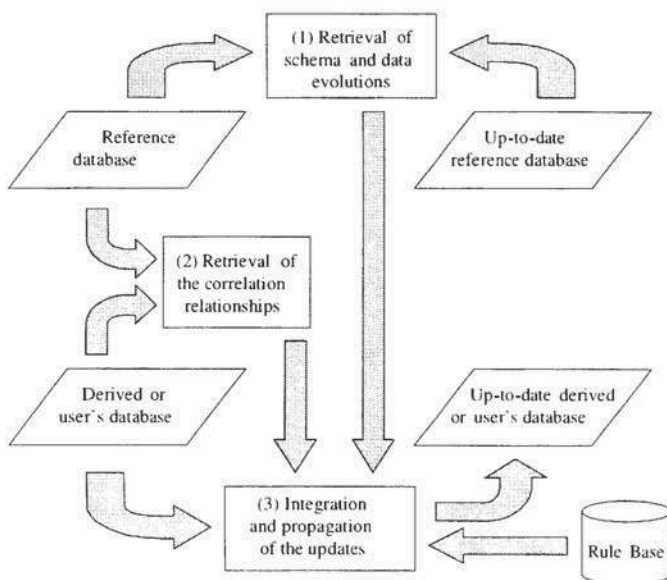


Figure 3 Structure of an updating mechanism for geographic databases [Badard 00].

Towards a more generic approach

Towards a more generic approach

The different studies previously presented have a common point: they use geographic information systems constructed on models which do not take into account the notion of evolution. This observation urges us to see the problem of update at the level of the representation of the data. We think that a reorganization of the geographic data view on updates can be a first step towards a generic solution of these various difficulties.

The originality of our approach in comparison with focuses studies lies in the fact that evolution of geographic information is taken into account concurrently with the modeling of systems and not afterwards. This approach focuses on the organization of geographic data in GIS to facilitate their updating and to preserve their consistency. Our approach is twofold:

- first, we propose a formalism for the representation of geographic data that takes their evolutions into account;
- second, we set up a reasoning method that relies on a model that allows the propagation of modifications and the preservation of consistency during the updating process.

This general frame is then put to use in the particular context of the update of a geographic database. More exactly, given an update information, the system integrates this update into the base thanks to the reasoning method that takes into account the particular formalism on which the base must be constructed.

The model

A quadruplet (D, R, M, C) represents our model where:

D denotes the domain of handled variables. These variables take their values in the set of all geographical objects stored in the base. We consider three types of objects:

- complex objects, which are composed of simple objects or complex objects; for example, a road is composed of road sections;
- simple objects: for example a road section or a building;

- geometrical primitives: points, lines and surfaces.

R denotes a set of relations which can exist between the objects. According to the nature of the knowledge they carry, we distinguish four main types:

- distance relationships: the distance may be qualitative (the house is far from the hospital) or quantitative (the house is 100m away from the hospital);
- topological relationships: they translate the notion of sharing of geometry which can exist between two objects such as adjacency and inclusion (two buildings sharing the same wall);
- composition relationships: a road is composed of several road sections;
- correspondence relationships; a complex object (eg. roundabout) corresponds to a simple object (eg. node).

M is a set of evolutions to be integrated in the database. It gathers old (to be modified) and new (already modified) versions of objects and the type of the evolution undergone by the objects.

C represents a sequence of constraints that allows on one hand to discover the conflicts and on the other hand to make a decision to solve conflicts. We consider two types of constraints:

- semantic constraints: all the constraints relating to the geographic domain and to the specifications of the base : example a lake should be horizontal.
- geometrical constraints: all the constraints relating to the geometry of objects : example, A and B are two objects; if A is included in B and B is included in A then geometry of A equals geometry of B.

The general structure

In our model we make the updates to the base level of the object representation. By "base level" we mean the level of geometric representation of simple objects. We call it the syntax level. Then, in the semantics level we study under which conditions these updates can or cannot be propagated to the other objects. Figure 5 supplies an illustration of the layered architecture of our model.

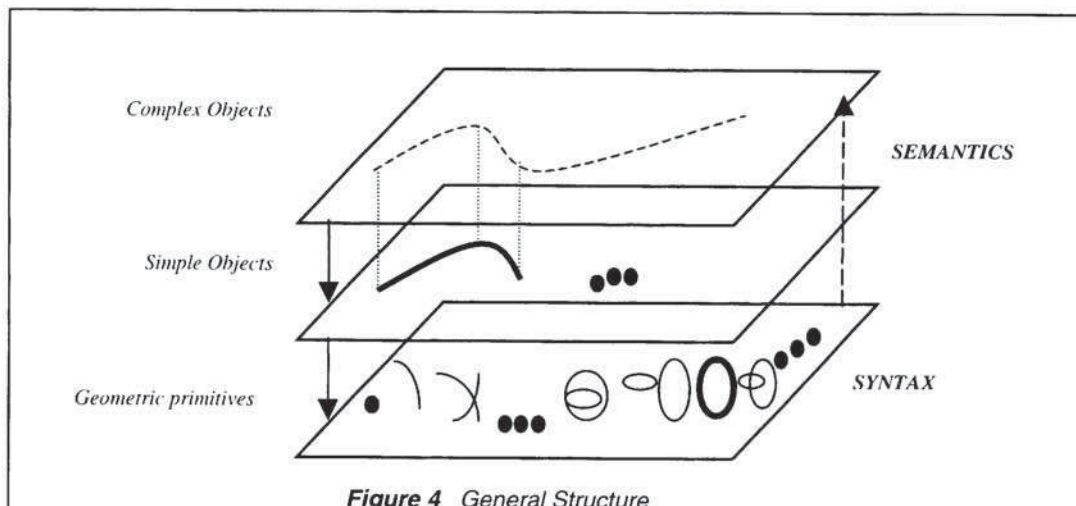


Figure 4 General Structure

The syntax level

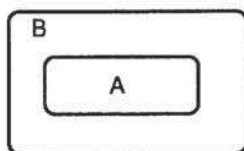
In this level, we establish the formalism for spatial object representation. Geometrical objects (primitives) are handled. This formalism must be sufficiently general to represent the spatial data and specific enough to answer the requirement of updating. To achieve it, geometrical primitives necessary for the representation of the objects in the database are determined. Relationships between objects are also determined and all the updating operations which can affect the base are defined. Finally the effects of each update operation on the relationships as well as on the entities are managed by the constraints. These constraints determine what objects and what relationships are selected, and how they are going to be modified.

The semantics level

The knowledge in the semantics level completes what we have at the syntax level. Indeed, the only information that we have at the syntax level are of geometrical order, and they are not sufficient for the detection or the resolution of a conflict.

Example

Surface(A): A is a surface;
Surface(B): B is a surface;
Delete(A): A is deleted;
Delete(B): B is deleted;
Include_B(A): A is included in B;
C=[Surface(A) and Surface(B) and Delete(B)]fi Delete(A)



B is deleted. According to constraint C, the destruction of object B implies automatically the destruction of object A. In that case, one goes up to the semantics level to have more knowledge on these two objects in order to make a decision as for the destruction or not of object A. If B is a wood and A is a glade, the destruction of A when B is destroyed seems logical. Let us suppose now that A is a building inside of a wood represented by object B. In that case, the destruction of A is not automatic and we can not take the decision to destroy A.

Conclusion and outlooks

In this paper we have just surveyed some of the key ideas and results of some research works which have addressed the problem of geographic databases updating. Together with these works, several research have been developed to make the reasoning on the spatiotemporal data easier. Interfacing these techniques with updating works is an interesting and important challenge. It is what we try to investigate in our research work.

Future works will consist in the establishment of the representation formalism and the set up of the reasoning method. Formalizing the relations between objects and making them explicit will allow for the discovering of all the conflicts which appear during the updating. Establishing the constraints will allow the system to propose solutions to these conflicts.

References

- [Badard 00] T. Badard. Propagation des mises à jour dans les bases données multi-représentation par analyse des changements géographiques. Thèse de doctorat, Université de Marne-la-Vallée, décembre 2000.
- [Bédard and al 97] Y. Bédard, Y. van Cheistein & G. Poupart : Actualisation des données à référence spatiale (volets échange et intégration), Centre de Recherche en Géomatique, Université Laval, Québec, Canada, 54 pages, 1997.
- [Bonani 98] L. Bonanni. Établissement de liens de corrélation dans un but de mise à jours des bases de données géographiques. Mémoire de DEA, Systèmes Intelligents, Université Paris IX Dauphine, laboratoire COGIT, IGN-SR. 29 septembre 1998.
- [Bréard 00] J.-Y. Bréard. Mise à jour en continu, Aspects techniques- Étude préalable. Rapport technique interne, Direction Technique, IGN, Mars 2000.
- [Egenhofer and al 92] M.J. Egenhofer, J. Herring. Categorizing Binary Topological Relationships between Regions, Lines and Point in Geographic Databases. Department of Survey Engineering, University of Maine, 1992.
- [Harrie and al 99] L. Harrie, A. Hellstrom. A case Study of Propagating Updates between Cartographic Data Sets. Proceedings/Actes, 19th International Cartographic Conference, 11th General Assembly of ICA, Ottawa, 1999.
- [Kilpelainen 95] Kilpelainen, T. Updating Multiple Representation Geodata Bases by Incremental Generalisation. Geo-Information-System, Jahrgang 8, Heft 4, Wichmann, pp.13-18, 1995.
- [Raynal 96] Raynal, L. Some elements for modelling updates in topographic databases. In the proceedings of GIS/LIS'96, Annual Exposition and conference. Denver, Colorado, USA, November 19-21, pp. 1223-1232, 1996.
- [Spéry 99] L. Spéry. Historique et mise à jour de données géographiques : application au cadastre français. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, octobre 1999.
- [Uitermark and al 98] H. Uitermark, P. Oosterom, N. Mars, M. Molenaar. Propagating Updates Corresponding Objects in a Multi-source Environment. Proceedings 8th International Symposium on Spatial Data Handling, pp.202-213, 1998.