Unifying Databases: Experiments with a Hydrographic Database

Sébastien Mustière[1], Adrien Cleach, Julien Fort[2]
[1] Institut Géographique National, COGIT Laboratory, Saint-Mandé France
sebastien.mustiere@ign.fr
[2] Ecole Navale IRENAV, Lanvéoc-Poulmic France

SUMMARY

This paper presents some research on the unification of different geographic databases. We describe the interest of multiple representations and databases unification, for data users as well as data managers. We also define a global framework for the unification of geographic databases: from the individual databases analysis to the definition and instantiation of a unified database. Then, we describe results from experiments led to unify the hydrographic themes of two actual DB from IGN-France with different scales. The teachings of these experiments concern the role of implicit information encountered in geographic databases, either inside one database or between two databases. These experiments also enlighten the need of formal information management tools.

KEYWORDS: Vector geographic databases, unification, integration, multiple-representation.

INTRODUCTION: UNIFICATION OF GEOGRAPHIC DATABASES

In order to reflect the diversity of points of view on the geographical world, many geographical databases exist to represent a same part of the world. These databases can either have different levels of detail, represent different objects, or represent the same objects but in different manners.

Unfortunately, most of the time these databases are relatively independent. But there exists a growing need for the unification of different databases into a single one making explicit the relations between them [Devogele, Parent and Spaccapietra 1998], possibly with the use of multiple representations [Vangenot, Parent and Spaccapietra 2002]. Such a unification would help to:

- Maintain the databases and propagate *updates*. For data managers, this is important to decrease the cost of updating data : updates may be integrated once and propagated, at least semi-automatically, to different representations of a same geographic phenomenon. For data users, unification could help to easily integrate updates coming from data producers, without loosing their own enrichment of data.
- Perform some *quality analysis*. One representation can be used to control another one or to identify inconsistencies [Egenhofer, Clementini and Di Felice 1994, Sheeren 2002]. For data managers, evaluation of quality is important to control and increase the quality of their data. For data users, evaluation of quality is important to assess the fitness for use of data.
- Increase the *potentiality of applications* using these databases. For data users, the use of data coming from different databases would be much easier if these data are well integrated. For data producers, the development of new databases including data coming from existing databases would be facilitated.

UNIFICATION PROCESS

In order to unify databases, we identify several main steps to be done, as illustrated in Figure 1. Before detailing and illustrating these steps in the next section with experiments made to create a unified hydrographic database, we briefly describe them hereafter:

- Independent databases are analyzed, in order to understand their content and prepare them before unification. This step requires the joint analysis of data, schema, and specifications.
- The links between the schemas are identified. In particular, one must identify which classes are suppose to represent the same objects, and how attributes of one database are related to attributes from the other database.
- A data schema, and related specifications, are defined according to the available data and to the user needs.
- At the data level, links are established and interpreted between data. A matching process must be used to identify corresponding objects between databases [Devogele 1997]. But a single matching is not enough at this step: matching links must be evaluated and inconsistencies must be detected and interpreted. One must differentiate between differences that are normal according to the specifications, differences that are due to updates, and differences that are due to errors in one of the databases [Sheeren 2002].
- Finally, the unified database can be instantiated. The defined schemas and specifications guide the transformation of data from single databases into data in the unified database.

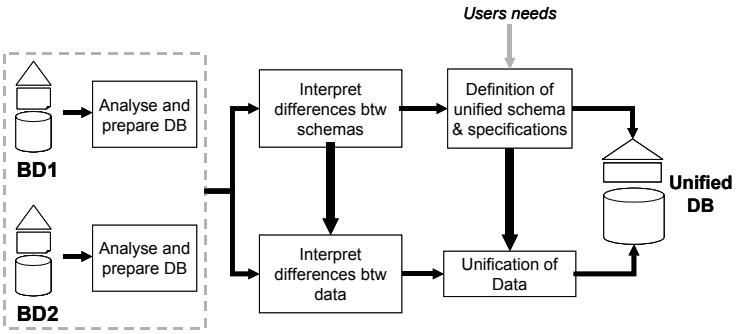

Figure 1: Unification process

A HYDROGRAPHIC DATABASE WITH MULTIPLE REPRESENTATIONS

Presentation of databases

Experiments have been conducted from two databases from the French National Mapping Agency (IGN): BDTopo and BDCarto (see Figure 2).

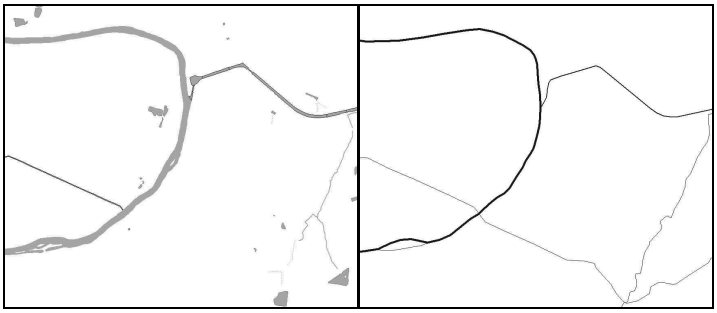

Figure 2: A view of hydrographic themes of BDTopo (left) and BDCarto (right)

BDTopo has a metric precision and is dedicated to description of topography. The river network may be described by either lines or surfaces. BDCarto has a decametric precision and is in particular dedicated to the description of networks. The river network is mainly described by lines. From the producer' point of view, unifying these databases is an important challenge, first of all to facilitate their management and their updates.

Analyzing and preparing independent databases

The first important task for unification is the independent analysis of individual databases. For example in the hydrographic themes of our databases, the class *hydrographic sections* of BDCarto includes the aqueducts, but the class *hydrographic sections* of BDTopo does not include them (they are in a class named *pipe* in the theme *distribution networks*). All these information, i.e. the precise definition of the elements of the schema, is only described in some textual documents: the specifications of the databases. This analysis of specifications, with an analysis of data and schemas, is a necessary task for handling the complexity of the data.

This analysis of specifications, data and schemas, reveals a lot of implicit information in the database. In order to prepare the unification, we reorganized and enriched the data schemas to explicitly represent these information. The main transformations on the schemas were:

- Detect and represent in the theme some external classes. Indeed, a lot of information concerning hydrography are outside the *hydrography* themes in the original schemas. For example, *ponds* are elements of a class *surfacic land use* in the *land use* theme of BDCarto.
- Add classes to better organize the schemas. We added classes to better organize existing classes through some hierarchies. For example the classes *man made elements* and *natural elements* were added and linked to existing classes by inheritance.
- Represent implicit relations between classes. This may concern classes within the considered hydrographic theme or with other themes. For example, limits of *land use surfaces* should exactly match *high water marks* when this is relevant: this relation was only implicitly described in specifications and has been added to the schema. We consider that the task of adding implicit information may be the most critical part of the unification work.

This task of analysis is a complex task, based on a joint analysis of data, schemas and specifications. In order to facilitate it, two main directions could be studied: the formalization of specifications [Mustière, Gesbert et Sheeren 2003], and the development of spatial data mining to automatically detect implicit relations in data.

Matching schemas

The next step is to match the data schemas, in order to identify how the elements of the model (classes, attributes and relations in an object-oriented paradigm) relate to each other. We defined in the previous step some new individual schemas, with a better organization and with the expression of hidden information. For schema matching, we modified the individual schemas again, in order to make them as close as possible. We then obtain schemas very rich and close, expressed in the same language with the same principles. Their matching is thus easier.

The matching is not only an identification of linked classes. The relations between classes must be precisely characterized. For example, the classes *land use surfaces* of BDCarto and *hydrographic surfaces* of BDTopo are linked. But only the elements of the class *land use surfaces* with the attribute *nature* equals to "*open water*" or "*swamp*" may match some elements of *hydrographic surfaces*. This is an important information that needs to be identified.

The automation, or at least semi-automation, of this task is highly difficult. If this is done interactively, the determination of matching classes is a long but quite direct task if, and only if, data schemas and specifications have been first precisely analyzed. Anyway, several issues still need some more researches:

- *Precise matching of attributes*. It is rarely direct to assess how attributes relate each other, without a long analysis of data and specifications. But this could be partly automated once the global links between classes have been defined, through an analysis of sample data.
- *Consideration of relations*. Indeed, few geographical databases really have explicit relations: these relations are only implicit through the geometry. Thus, the matching of relations in schemas did not receive yet a lot of attention.
- *Languages of description of relations between schemas*. We used the "inter-schema assertions" language described in [Devogele, Parent and Spaccapietra 1998]. This language has been suitable for this task, but we felt a lack of formalization for expressing some geometric relations between classes.

Defining a unified data schema

Once schemas have been analyzed and linked, the next step is the definition of a data schema for the unified database. This is a task, as any modeling task, that should pay attention to the intended use of the database. In our case, the intended use was to facilitate updates. One constraint was to be able to re-derive the original databases when needed. Thus we should not afford to loose any information. We thus decided to create a schema with multiple representations of geographic phenomena. An extract of this schema, focusing on *water marks*, their multiple representation, and their relation with other classes, is shown in figure 3.

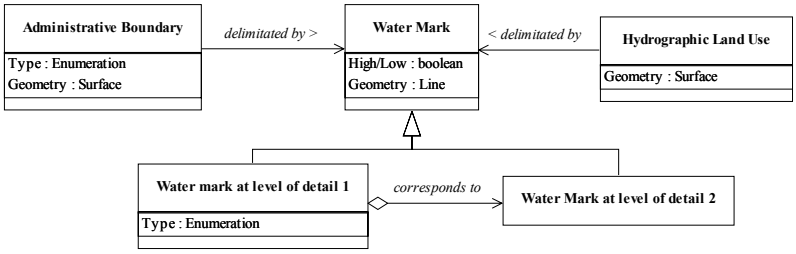

Figure 3: extract of the unified schema (UML language)

In terms of methodology, the creation of this schema has not been straightforward. We first combined the two detailed individual schemas. We obtained a very rich schema, expressing all the possible links. But this schema was too complex to be intuitively handled. We thus transformed this schema into a much simpler one. Anyway, we believe that the step of building an over-complex schema before simplifying it was necessary to handle the complexity of geographic databases.

The main difficulty of this task has been to formally express how the schema of the unified database can be related to the individual schemas. Once again, we used the "inter-schema assertions" language, but we noticed that this language was more suitable for expressing links between existing data, than for expressing how a schema can be derived from another one.

Matching data, interpreting inconsistencies and instantiating the unified database

Matching at the schema level provides information about which objects of the two databases may be matched together. The instantiation of the unified schema necessitates to identify actual matching data. For the time being, we did only experiment one automatic matching process between linear networks. We used an algorithm developed for road networks [Devogele 1997]. The results are satisfying in some cases. But the results are rarely good, even if the used algorithm is efficient. Indeed, we did not pay attention to the preparation of data before matching, and thus encountered many problems. The manipulated data were not well organized in a linear network: some nodes were missing, and some hydrographic area did cut the linear network, without insuring the connectivity with additional lines. We believe anyway that

automated matching with very good results is possible, even if the task is too complex to expect a fully error-free automated process.

The task of matching is not enough for unification. Once data have been matched, it is necessary to analyze the matching, and particularly to detect and interpret inconstancies [Sheeren 2002]. For example, once we have matched the data, we looked for inconstancies like in the following case: one river with an attribute equals to "*permanent*" is matched in the other database with one or several rivers with an attribute equals to "*temporary*". We did not try yet to explain this inconsistencies: for example, it may be possible to analyze surrounding rivers and terrain in order to interpret which data may be wrong. We did not either try to detect inconsistencies between the geometric representations of objects. Finally, we did not instantiate yet the unified database, and especially the links between the different representations of data. These tasks are an undergoing work.

CONCLUSION

We presented a global framework for the unification of geographical databases. This framework is in five steps. The first step is the analysis of the individual database, and is based on the joint analysis of data, specifications and schema. The second and third steps are at the level of the schemas: determination and precise description of the relations between individual schemas; development of a unified data schema and the associated specifications. The two last steps are at the data level: matching of the data and interpretation of the differences; instantiation of the unified database.

In this paper we presented experiments made on hydrographic data from two actual databases. Databases have been analyzed, schema have been matched, and a unified schema has been defined. First experiments on data matching have been made. The main teachings of this study are:

- During our experiments we built a lot of different complex schemas. Because of the difficulty of the process, the process should follow a formal information management framework. We also need to express a lot of information in formal models.
- In our context, the most challenging particularity of geographic data is the importance of implicit information, hidden through geometric relations, inside and between databases.

BIBLIOGRAPHY

Devogele T., 1997. Processus d'Intégration et d'Appariement de Bases de Données Géographiques - Application à une bases de données routières multi-échelles. PhD Thesis, Université Marne-la-Vallée.

Devogele T., Parent C. and Spaccapietra S., 1998. On spatial database integration. In International Journal of Geographical Information Science, vol.12, n.4, 335-352.

Egenhofer M.J., Clementini E. and Di Felice P. 1994. Evaluating inconsistencies among multiple representations. In Proceedings of the Sixth International Symposium on Spatial Data Handling (SDH'94), Edinburgh, Scotland, 901-920.

Mustière S., Gesbert N., and Sheeren D., 2003. A formal model for the specifications of geographic databases. In Proceedings of 2nd workshop on Semantic Processing of Spatial Data, Mexico, S. Levachkine, J. Serra and M. Egenhofer (eds). 152-159.

Sheeren D., 2002. L'appariement pour la constitution de base de données géographiques multi-résolutions - vers une interprétation des différences de représentation. In Revue Internationale de Géomatique, vol.12, n.2, 151-168.

Vangenot C., Parent C., Spaccapietra S. 2002. Modeling and manipulating mutliple representations of spatial data. In proc. of 10th International Symposium on Spatial Data Handling, SDH 2002, Ottawa, Canada, pp.81-93.