FORMALISING THE GEOGRAPHIC DATABASE GENERALISATION PROCESS BY MEANS OF A CONFLICTS/OPERATIONS GRAPH

Dominique Han-Sze-Chuen, Sébastien Mustière , Bernard Moulin

Computer Science Department and Geomatics Research Centre, Laval University,
G1K 7P4, Ste-Foy, Québec, Canada
dominiquehan@yahoo.com, {bernard.moulin, sebastien.mustiere}@ift.ulaval.ca

KEY WORDS: Databases, Generalisation, GIS, Modelling.

ABSTRACT:

Database generalisation is a process which aims at producing a geographic database from an initial geographic database in order to satisfy new needs. Starting from the study of an actual database generalisation process developed at the Ministère des Ressources Naturelles of Québec (MRNQ), our intention was to formalise the cartographic knowledge contained in the process to make it reusable. MRNQ's process has the form of a workflow which contains a list of actions applied to object classes. Since we wanted to extract generic knowledge, we introduced the notions of *problem* and *operation*: our assumption is that the cartographic knowledge formalised using these two notions is generic enough to be applied to other cases of generalisation. Thus, we propose a causal graph which associates a set of problems to a set of operations. The utility of such a model is that it offers an explanation for each actions contained in the workflow and allows an anticipation of future problems when creating a new workflow.

RÉSUMÉ:

La généralisation de base de données est un processus qui vise à produire une base de données géographiques, à partir d'une base de données géographiques initiale en vue de répondre à de nouveaux besoins. En nous appuyant sur l'étude d'un cas réel de généralisation développé au Ministère des Ressources Naturelles du Québec (MRNQ), notre objectif a été de formaliser les connaissances cartographiques contenues dans ce processus afin de pouvoir les réutiliser par la suite. Le processus du MRNQ se présente sous la forme d'une chaîne de production contenant une liste d'actions qui s'appliquent sur des classes d'objets. Du fait que nous voulions extraire des connaissances génériques, nous avons introduit les notions de *problème* et d'*opération* : notre hypothèse est que les connaissances cartographiques, formalisées à l'aide de ces deux notions, sont assez génériques pour être utiles à d'autres cas de généralisation. Nous proposons ainsi un graphe causal qui associe un ensemble de problèmes à un ensemble d'opérations. L'intérêt d'un tel modèle est qu'il fournit une explication pour chaque action de la chaîne de production et qu'il permet d'anticiper d'éventuels problèmes lors de la création d'une nouvelle chaîne de production.

1. INTRODUCTION

Geographic data generalisation is a key process to produce geographic data at different levels of abstraction, either maps or geographic databases (GDB) [Brassel and Weibel 88]. In this paper, we focus on the process of creating a GDB from another GDB whose content is too detailed when considering the product to be generated. This process is sometimes named *model generalisation* [Weibel and Dutton 99] in order to distinguish it from *cartographic generalisation* that aims at producing graphic maps.

The creation of different GDB at different levels of detail is useful for different purposes: to reduce the size of a GDB, to allow spatial analysis for various purposes and at different scales, and to create pivot databases dedicated to the derivation of different thematic products.

Due to the complexity of the process, there still exists no complete automation of geographic data generalisation [Joao 98; Weibel and Dutton 99]. To date, no GIS provides the means to automatically generalise data. Few of them provide a toolbox of generalisation algorithms, but the cartographer's intervention is still needed at two levels. First, s/he must define how to use the toolbox algorithms according to the target product, either to develop a global process or to interactively select algorithms during the generalisation process. Then, s/he must perform manual operations, either to correct the results provided by the algorithms or to manually perform operations for which no algorithm is provided.

Our work starts from the study of an actual generalisation process developed at the Ministère des Ressources Naturelles (MRN) of Quebec. Several meetings with the cartographers of the MRN and several documents describing this particular generalisation process enable us to collect a large amount of cartographic knowledge about model generalisation. The purpose of our work presented in this paper is to study how to formalise the knowledge collected from this particular generalisation process in order to make it reusable for others cases of generalisation.

In the second section of this paper, the MRN's generalisation process is described. Then, in the third section, a model of representation of cartographic knowledge in generalisation is proposed by means of a causal graph. Finally, in the fourth section, we discuss the advantages and limits of the proposed model.

2. ANALYSIS OF AN ACTUAL GENERALISATION PROCESS

2.1 The new database context production

Our study dealt with an actual generalisation process developed at the *Ministère des Ressources Naturelles (MRN)* of Quebec. This process aims at producing a 1/100,000 scale topographic database (namely the BDAT) from a 1/20,000 scale topographic database (namely the BDTQ). This work is supported by the Intergraph's DynaGen software, which is a platform dedicated to model generalisation.

This workflow have several characteristics. First, a good quality of generalisation is needed because the BDAT is a pivot database that will be used for several purposes. Second, because of cost and time constraints, the MRN's generalisation process has been automated as much as possible. Then, due to the fact that the production of the BDAT will be subcontracted to private companies, the generalisation process is very detailed in order to insure a good and homogeneous quality of the results.

Consequently, in a reverse engineering point of view, this workflow is a great opportunity to collect a large amount of cartographic knowledge about the generalisation process.

2.2 The workflow

In order to specify how to produce the new intended database (called the final database in this paper), the MRN's cartographers developed a workflow decomposing the generalisation process into a sequence of basic steps associating certain groups of objects to specific actions to be applied on them (Figure 1).

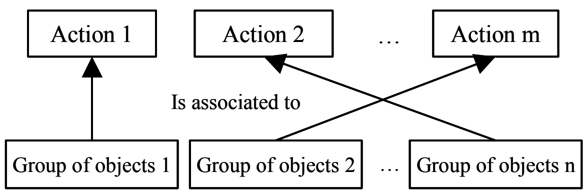

Figure 1: the workflow links groups of objects to actions

A group of objects which must be generalised is defined in the workflow either by a full class of objects or by a condition on the objects of a class (e.g. all the objects of the class *2-D rivers* whose *minimum width* is less than *100 m*; or all the objects of the class *lake* whose attribute value *area* is under *30,000 m2*). An action which is performed is either manual or automated (e.g. select all isolated lakes by means of a query, visually check the data at the junction of two datasheets, apply the algorithm *area to line* with parameters *5 and 0.5,* apply the algorithm *Douglas and Peucker* with the parameter *0.05* [Douglas and Peucker 73]).

It must be noticed that many successive versions of this workflow have been developed and tested by specialists from MRN during several months, before getting an adequate process. Another important point is that the workflow still contains manual operations.

2.3 Purpose of the work

We wondered if the cartographic knowledge included in the workflow could be reused in another generalisation process using other data, with another generalisation software in order to produce another product.

It is important to notice that this workflow is dedicated to a particular GDB creation, from a particular GDB, with a particular tool, in a particular organisational context. Even if the workflow is well documented, it may be over-detailed to be easily reusable when developing a new generalisation process. Anyway, our assumption is that this workflow contains some generic knowledge about which actions should be done in various cases that can be encountered during different generalisation processes.

The purpose of our work is then to:

- *Organise* the knowledge involved in the workflow in order to emphasise the most relevant information potentially useful during the development of a new generalisation process.
- *Extract* the most generic (and then reusable) knowledge from the workflow.
- *Reformulate* the knowledge involved in the workflow in order to make it easily adaptable to a new generalisation process.

3. OUR MODEL FORMALISING THE DATABASE GENERALISATION

3.1 Extraction of the generalisation actions

If the workflow contains many generalisation operations that may be relatively generic, it also takes into account specific problems due to the characteristics of the database and the generalisation software used. The solutions to these problems are specific actions such as changing certain object attributes in the database or simplifying objects before changing their dimension to improve the efficiency of reduction algorithms. In fact, these specific actions can be thought of as a pre-processing which aims at preparing the data to be generalised.

Here is a list of the main reasons justifying the performance of specific actions:

- The initial database may contain some minor errors that could be emphasised during the generalisation process. These errors have to be corrected.
- The initial database model is not adapted. Some modelling choices made in the BDTQ are more adapted to directly draw the data on a map rather than automatically analysing and transforming them.
- Some generalisation algorithms are not efficient in some particular configurations. Actions must then be done to prevent problems in these cases or to to correct them afterwards.
- There exists no algorithm to make certain intended operations. Series of operations are then made to overcome this deficiency.

Hence we must distinguish, first, the actions which one performed to prepare the data and are very specific to the initial data and, second, the actions performed to generalise the data, which means transforming the data in order to respect the specification of the final database. Since our goal is to formalise generic knowledge, our model will only take into account generalisation actions.

3.2 Reformulation of the actions of the workflow in a problem / operation formalism

The actions described in the workflow directly link some object classes of the initial database to algorithms used to transform them (e.g. all objects of the class *River* must be transformed by the algorithm *Simplification* with a parameter *p*). The drawback of this approach is that these object classes are specific to the BDTQ. Hence, in another case of generalisation, the initial database may contain different classes of objects. Since the interest of formalising the knowledge contained in the workflow is to make it reusable, associating a class to an action is not a good option because such an association will be relevant only to MRN's particular generalisation process.

In order to extract generic knowledge from MRN's workflow, we use the notions of *problem* and *operation*. Our assumption is that certain generic problems must be solved with certain generic operations whatever the case of generalisation. The generic part of knowledge relies on the links between problems and operations, while the specific part of knowledge relies in the links between, on the one hand, objects and the problems they have, and on the other hand, operations and the algorithms used to perform them. Thus, as described in Figure 2, our model will consider the generalisation process as a set of *problems* that must be solved by applying certain *operations* [Armstrong 91; McMaster 91; Shea 91].

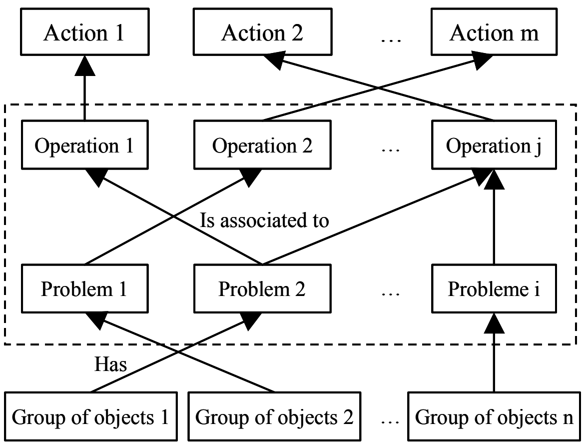

Figure 2: Introduction of the notions of *problem* and *operation* in order to extract generic knowledge

A problem results from the fact that a constraint on the final product is not respected [Beard 91; Weibel 96; Ruas 00]. It must be noticed that the problems are not explicitly described in the workflow. The intentions of the cartographers need to be reformulated in a way to describe the problems and the solutions. For instance, if a group of rivers is simplified with *Douglas and Peucker* algorithm, we reformulate this fact to emphasise the reason why this has been done. The proposed reformulation using the notion of problem and operations is: there are in the initial database some objects (the rivers) having a *geometric shape which is too detailed* to be in the final database; and the solution proposed by the cartographers to solve this problem is to *simplify* them (using Douglas and Peucker algorithm). By this way, it is possible to note the problem *geometric shape too detailed* and its solution *simplification*. This kind of knowledge is reusable because in another case of generalisation we suppose that there will still have groups of objects with a geometric shape too detailed and that the solution will still be valid.

In the next sections we will detail the types of problems that we identified, the types of operations selected to solve them, and finally we will associate the problems to the operations.

3.3 Problems and operations

In order to reformulate the workflow into a set of problems linked to a set of operations, one must first identify generic problems and generic operations.

First, the workflow enables us to distinguish two types of problems: problems on one object and problems on one group of objects. More precisely, we decomposed problems on one object into problems related to the object itself and problems related to the relations that this object has with other objects. In particular, we noticed that numerous problems did imply a particular relation: the 'support relation' between one object and another object that must exist to ensure the coherence of the geographic space (e.g. if a house is supported by an island, the house must be removed if the island is removed).

Figure 3 presents a more detailed classification of the problems found in MRN's workflow.

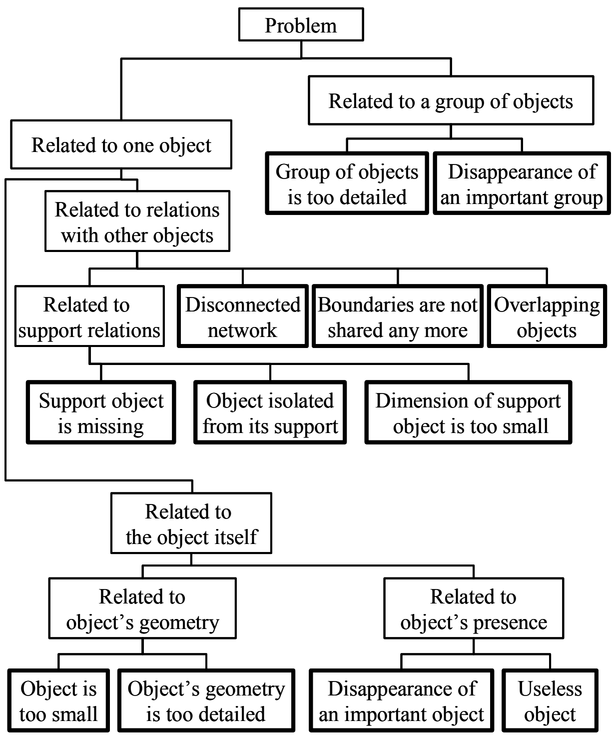

Figure 3: Classification of problems found in the workflow

Second, we classified the operations made in the workflow. Many classifications of generalisation operations do exist (see for example [McMaster and Shea 92; Peng and Tempfli 96]). For our purpose, we used the classification described in Figure 4. One can notice that the classification follows the same principle than the one used to classify the problems. Operations are classified into two categories: operations on an object and operations on a group of objects. These categories are further divided into subcategories which are associated to typical generalisation operations: simplification, reduction, deletion, etc.

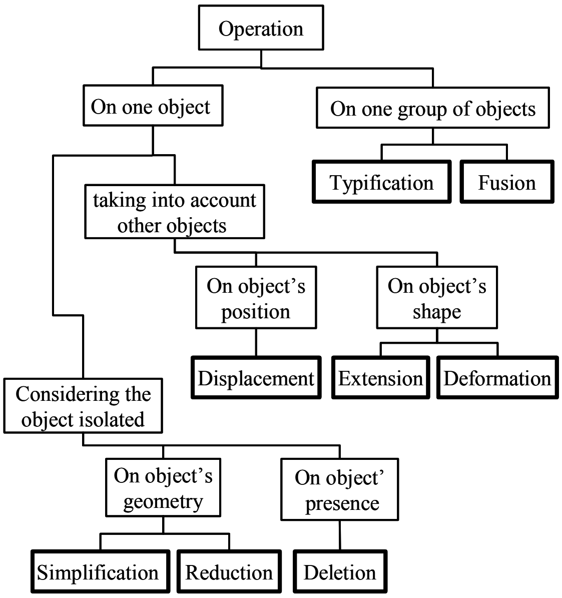

Figure 4: Classification of operations made in the workflow

3.4 A causal graph linking problems and operations

The analysis of the workflow and several interviews with cartographers allowed us to develop a causal graph to represent generic knowledge found in the workflow (Figure 5, on the right). The graph associates problems to be solved (square boxes) to the generalisation operations (round boxes) that must be applied to solve these problems. It contains knowledge about which operations allow to solve a given problem (dash lines), and about which derived problem may arise from the application of a given operation (continuous lines). The graph also distinguishes initial problems that may appear at the beginning of a generalisation process (in bold) from derived problems that may result from the application of certain operations. Each link in the graph can be described by a form that explains more in detail why this link does exist and provides graphic examples.

This graph is generic in the sense that it does not depend on a particular generalisation process. This makes this graph very useful for several purposes, as it is explained in the next section.

4. UTILITY AND LIMITS OF THE MODEL

4.1 Utility of the causal graph

The first utility of the graph appears when one wants to analyse a special process such as the MRN's workflow. One can do that by associating each step of the process to the path of the graph it corresponds to.

On the one hand, this allows to distinguish the generic and context-related steps of the process. Indeed, the steps of the process that do not correspond to any path in the causal graph are not generalisation steps. We thus can distinguish steps that are actual generalisation operations and steps that are introduced to overcome the system limits (e.g. inefficient algorithms) or to deal with the initial GDB defects. On the other hand, the graph creates an organisational memory. Indeed, as we explained before, the graph is a way to extract, organise, and reformulate the generic knowledge involved in the generalisation process. In addition, the graph explains *why* these steps have been chosen (i.e. to solve which conflict), which is not stored in the raw process description. This is useful if the process has to be refined in the future or adapted to a new product.

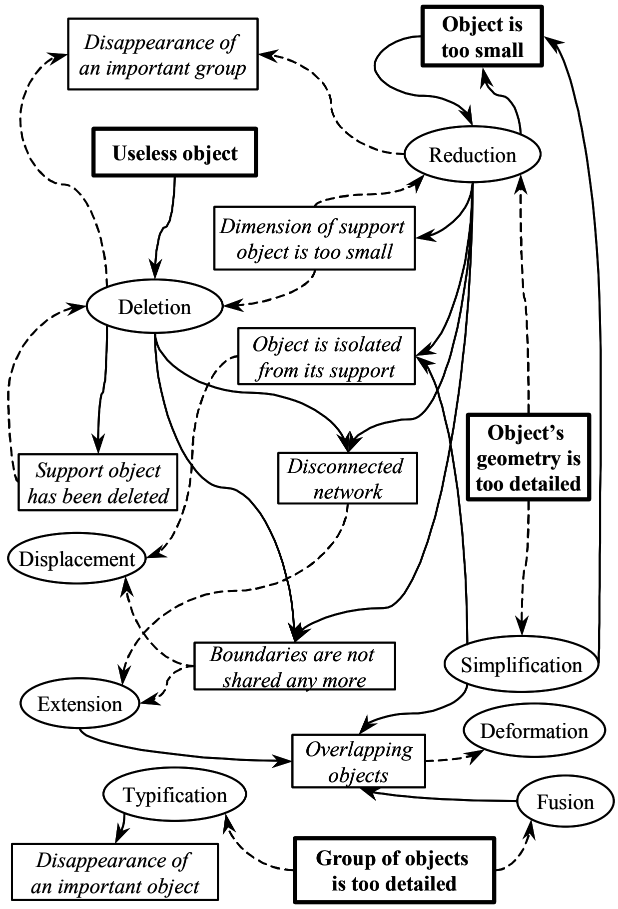

Figure 5: Causal graph linking problems and operations

The second utility of the graph is to help cartographers to define a new process, when there is a need for a new product. Indeed, thanks to the graph, the workflow designer is able to identify the derived problems that may result from the application of a given operator. Consequently, the number of trials before getting a satisfactory process will certainly be reduced.

Another utility of the graph is to pave the way to a more computer-assisted generalisation process. As proposed in [Weibel 91], one can imagine generalisation platforms that contain both interactive operations, automated algorithms and knowledge bases used to guide the process. Such a platform could use the causal graph as part of its knowledge base in order to propose to the cartographers a set of operations to apply in certain conditions. Of course, this would require a platform containing some tools to help the cartographers to specify their needs (constraints on the final product), some efficient algorithms to automate the generalisation operations, and some efficient measures to automatically detect the problems due to constraints violations.

4.2 Limits of the model

One may argue that it is impossible to formalise all the generalisation knowledge in the form of a set of rules "if Condition then Operation", because there exist too many spatial configurations and too many possible generalisation solutions [Beard 91].

First, one must notice that this does not contradict the fact that the formalism "if Problem then Operation" can be efficient to model the knowledge involved in a particular process, in order to make it reusable.

Second, let us remark that our model only takes into account the generalisation process to produce databases, which is less complex than cartographic generalisation aiming at producing maps, because no legibility constraint appear. The causal graph may then encompass most of the cases encountered during database generalisation.

Finally, we believe that the causal graph presented in this paper is quite generic because of its degree of generality. For example, the problem "useless object" is not decomposed into all the cases in which an object can be useless. If this should be done, the graph may become very complex. The main drawback of this genericity is that this model cannot be implemented directly to automate the generalisation process. For example, the problem named “useless objects” should be refined into a precise specification which takes into account the characteristics of the final product (e.g. power lines are useless, narrow rivers are useless, small lakes except the isolated one are useless, etc.).

5. CONCLUSION

The analysis of an actual generalisation process is a good way to acquire knowledge about the generalisation process, but one must be able to differentiate specific knowledge from generic knowledge that can be reused. Our analysis of the workflow describing MRN’s generalisation process enabled us to obtain an initial model of the database generalisation process.

This model, represented by a causal graph, is a set of problems associated to operations. The introduction of problem and operation notions is a way to formalise generic knowledge that can be reused when dealing with others cases of generalisation. Indeed, because our model make explicit the problems underlying the workflow, it can used to anticipate the problems of a new workflow.

The proposed model is a good way to explain which generalisation operations should be done to solve certain problems. In order to go further, some work should be done to explicitly represent knowledge concerning how to sequence the generalisation process: when an object has several problems, which problem must be solved first? When several objects have problems, which object must be processed first?

ACKNOWLEDGEMENT

We would like to gratefully thank MRN's cartographers who have been available to explain us their workflow and to answer our questions. Thanks to the network of centres of excellence, GEOIDE, to National Defence and to Ministère des Ressources Naturelles du Québec for supporting this project.

REFERENCES

Armstrong M.P. 1991. Knowledge classification and organization. In *Map generalization: Making rules for knowledge representation,* B.P. Buttenfield and R.B. McMaster, (Eds), Longman Scientific & Technical, pp. 86-102.

Beard K. 1991. Constraints on rule formation. *Map Generalization : Making Rules for Knowledge Representation*, B.P. Buttenfield et R.B. McMaster (eds), Longman Scientific & Technical, pp.32-58.

Brassel K. et Weibel R. 1988. A review and conceptual framework of automated map generalization. *Int. Journal of Geographical Information Systems,* vol.2, no.3, pp.229-244.

Douglas D.H. et Peucker T.K. 1973. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or it’s Caricature. *The Canadian Cartographer journal*, 10(2), pp.112-122.

João E.M. 1998. *Causes and Consequences of Map Generalisation*. Taylor and Francis: London.

McMaster R.B. 1991. Conceptual frameworks for geographical knowledge. In *Map generalization: Making rules for knowledge representation,* B.P. Buttenfield and R.B. McMaster (Eds), Longman Scientific & Technical, pp. 21-39.

McMaster R.B. and Shea K.S. 1992. *Generalization in Digital Cartography*. Assoc. of American Geographers, Washington.

Peng W. and Tempfli K. 1996. An Object-Oriented Design for Automated Database Generalization. In *Proc. of the 7th Int. Symposium on Spatial Data Handling*, Delft, sec.4b, pp.15-29.

Ruas A. 2000. The role of meso objects for generalisation. In *Proc. of the 9th International Symposium on Spatial Data Handling,* Pékin, Chine, sec.3b.

Shea K.S. 1991. Design considerations for an artificially intelligent system. In *Map generalization: Making rules for knowledge representation,* B.P. Buttenfield and R.B. McMaster, (Eds), Longman Scientific and Technical, pp. 3-20.

Weibel R. 1991. Amplified intelligence and rule-based systems. In *Map Generalization, Making rules for Knowledge Representation,* B.P. Buttenfield et R.B. McMaster (eds), Longman Scientific & Technical, pp. 86-102.

Weibel R. 1996. A Typology of Constraints to Line Simplification. *Proc. of 7th Int. Symposium on Spatial Data Handling*, Delft, Vol.2, sec.9a, p.1-14.

Weibel R. and Dutton G. 1999. Generalising spatial data and dealing with multiple representations. In *Geographical Information Systems: Principles and Applications*, *2nd edition*, Maguire D.J., Goodchild M.F., Rhind D.W. and Longley P. (eds), Wiley, New York, Vol.1, pp.125-155.