DATABASES INTEGRATION FOR SUPPORTING THE FUTURE PRODUCTION OF IGN BELGIUM GENERALISED MAPS

Anne Féchir, Jan De Waele
Institut Géographique National – Nationaal Geografisch Instituut,
Cartography Department,
Abdij Ter Kameren 13, 1000 Brussels, Belgium -
afe@ngi.be, jdw@ngi.be

KEY WORDS : automatic generalisation, updates propagation, database integration, data conflicts, feature matching.

ABSTRACT :

In the near future, IGN Belgium will have to produce updated paper maps from a centralised large-scale geographic database. Two options were investigated in a production environment: automatic generalisation and updates propagation. This paper describes the first thoughts and results regarding the second option. It first considers the databases integration on a conceptual level, then goes deeply into the data synchronisation problem and finally describes a home-made algorithm that makes the link between the corresponding features in the different databases.

1. THE CONTEXT

IGN Belgium produces data from scale 1:10 000 to scale 1:250 000. The reference data at scale 1:10 000, once finished, will have taken 15 years to complete. The data at scale 1:50 000 edition 1 required 8 years of production.
The classic problem of different production cycles forced IGN Belgium to derive some of the generalised data from another source than the digital 1:10 000 data (Figure 1.). Half of the 1:50 000 scale sheets was based on the old 1:25 000 paper maps that were scanned, generalised and updated at great cost on the field. Most of the second edition 1:50 000 data, currently under process, is produced by collecting scale specific update information.

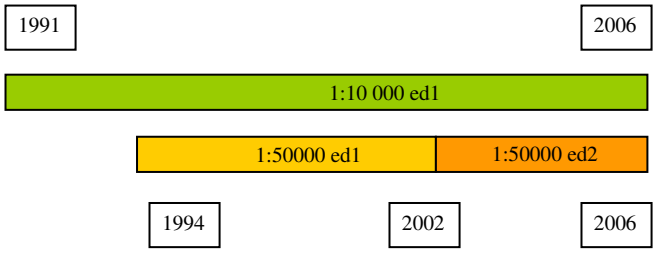

Figure 1. Production cycles

In 2001, a five years strategic plan with five main projects was defined. Among them, the SGISR Project (Seamless Geographic Information System of Reference) aims at the creation of a seamless GIS to manage topo-geographic reference data and to prepare IGN Belgium to contribute to the National and European Spatial Data Infrastructures initiatives.
The SGISR Project consists in a re-engineering process of the different production workflows of the topographic databases at 1:10 000 (3D-line and Top10v-GIS) and 1:50 000 (Top50v-GIS) in order to implement a unified maintenance of the data.
Six associated projects have been defined (Figure 2.).
As maps at different scales still have to be produced, a Generalisation Project is part of them.

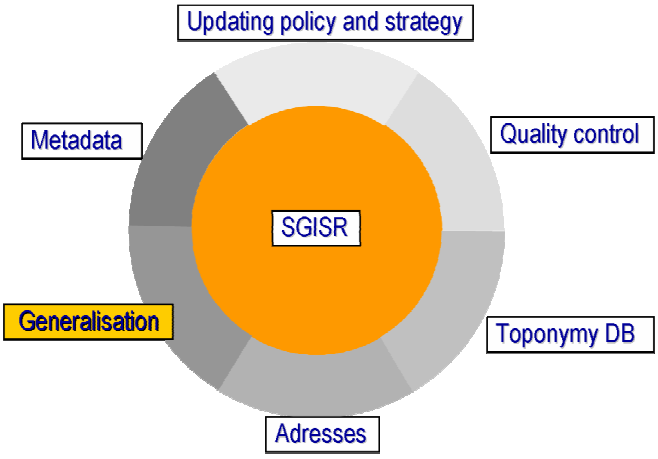

Figure 2. SGISR Project

2. UPDATES PROPAGATION VS. AUTOMATIC GENERALISATION

The main goal of the Generalisation Project is to produce, as automatically as possible, updated topographic maps at different scales from the reference data without using scale specific updating procedures. Two options were investigated: automatic generalisation and updates propagation (Ruas, 2002).

At this moment, full automatic generalisation is not yet possible for all the features represented on our maps. It is rather a long-term goal to achieve.
On the other hand, the updates propagation can be decomposed in several processes among which some can be automated in a quite short term. Updates propagation is also meaningful at IGN Belgium because, in 2007, both a reference dataset and derived datasets at scales 1:50 000 and 1: 100 000 will be available. Maintaining these generalised datasets can also be useful for certain applications.

In both cases, the aim is to develop a process that runs as automatically as possible.

Concerning the updates propagation, it should be possible to automate, in a short term, the selection of relevant updates in the reference dataset and their transfer or the mark up of the modified features in the derived datasets. The automatic generalisation of the updates and their integration in the derived dataset, maintaining the topological relationships and managing the neighbouring effects around the updated features will be investigated later.

However, automatic generalisation will stay the short term priority for features like the ordinary buildings. These features are very changing in Belgium and, most of the time, because they are typified, only a buildings group to buildings group relationship can be found between the corresponding features at different scales. This makes the updates propagation less straightforward. For those features, we would rather go for an automatic generalisation using the Agent technology (AGENT, 2000).

3. DATABASES INTEGRATION

« A major reason for a national mapping agency to investigate and implement a multi-representation database is the possibility of propagating updates between the scales, which is also called "incremental generalisation" » (Anders and Bobrich, 2004).
In order to propagate the updates from the reference data to the generalised data, we need to completely integrate the different databases, i.e. to merge the data from the different databases into a unified model and also to link the features from the different representations that represent the same real world phenomena, thus creating a multi-representation database (Devogele, 1997). This integration will allow us to automatically mark up the modified features in the generalised data. In addition, the integration has other advantages. It allows a quality control of the generalised data and data consistencies checks. We also hope to take advantage of this integration to characterise the generalisation process that was used to produce the derived dataset, in order to help us in the generalisation automation.
In this paper, we will concentrate on the 1:10 000 and 1:50 000 databases integration. Some of the mentioned problems probably apply also to the 1:50 000 and 1:100 000 databases integration.

3.1. Conflicts detection between the databases

As the databases at scale 1:10 000 and 1:50 000 have been produced independently and for different applications (the 1:50 000 has a military purpose), their original data models were quite different from each other.
In 1999, a common distribution structure for the two sets of data was built. Most of the definition conflicts were solved through a hierarchical coding structure.

ST210 : Hospital (1:50 000)
ST211 : Non-university hospital (1:10 000)
ST212 : University hospital (1:10 000)

When no such relation could be found between the objects definition, two different codes were just coexisting.

When the data model for the reference data in the new SGISR was built, most of the obvious 1:50 000 particularities were taken into account. Then the two models were examined more thoroughly through academic collaboration. The objective of this study was to adapt the 1: 50 000 data model in order to facilitate its integration with the new reference data model .
Using (Devogele, 1997) nomenclature, the following conflicts have been identified. The word conflict must be understood as a difference that makes the integration between the databases more difficult. Of course, most of these conflicts were deliberately created by the generalisation process. These conflicts must not disappear, they just have to be solved to allow the databases integration.
One example of each conflict found is described.

Heterogeneity conflict

Geometric metadata conflict :

- Resolution conflict : the 1:10 000 resolution is higher than the 1:50 000 resolution
- Exactitude conflict : the 1:10 000 exactitude is higher than the 1:50 000 resolution

Class definition conflict

Classification conflict:

- Grouping conflict: the 1:10 000 data distinguishes between simple and double lanes railway lines, as the 1:50 000 data identifies simple and multiple lanes railway lines.
- Resolution conflict: the water point features in the 1:50 000 data match the three following features in the reference data: source, fountain and well.
- Data/Metadata conflict: the embankments are road attributes in the 1:50 000 data and objects in the 1:10 000 data.

Specification criteria conflict:

- Selection criteria conflict: rivers less than 100 meters long and connected at only one end to the network are not present in the 1:50 000 databases.
- Decomposition criteria conflict: roundabouts are only collected for the 1:50 000 if there are legible at that scale.

Fragmentation conflict:

- Granularity conflict: A road is segmented when one of its attributes changes locally on a sufficient length. The limit is 200 metres for the 1:50 000 data and 50 metres for the 1:10 000 data.
- Decomposition conflict: dual carriageways are represented by two lines in the 1:10 000 data. They are collapsed into one line in the 1:50 000 data.

Structure conflict

Classic structure conflict: the embankments are road attributes in the 1:50 000 data and objects in the 1:10 000 data.
Information storage conflict: the area separating two lanes of a dual carriageway is identified according to what it is covered with (grass, concrete…) in the 1:10 000 data. In the 1:50 000 data, the existence of this separation is an attribute of the road ("with median").

Semantic and geometric description conflict

N-ary description conflict between attributes : The 1:10 000 roads coating can be "Solid road surface" or "Gravel" while the 1:50 000 road coating is described using the NATO categories.

Geometric description conflict: Churches are polygon features in the 1: 10 000 data and point features in the 1: 50 000 data.

Data conflict

Data conflicts occur when corresponding attributes have different values due to errors, different sources, different update dates or generalisation operations like typification (ordinary buildings for example).

Most of these conflicts can be quite easily solved. The last one seems to be more difficult to handle.
The case of the generalised ordinary buildings was mentioned above. The next section deals with the problem of update date difference.

3.2. The first synchronisation of the different databases

An ideal situation would be to have data at different scales produced at the same time (or in any case having the same "age") and then an updating process of the large-scale data that would produce an easy to identify set of updates to propagate into the generalised database.
In reality, a few years generally separate the second edition 1:50 000 data from the first edition 1:10 000 data. The first synchronisation of the different databases will not be an easy process.
Let us consider the different situations:

1. Generalised data is older than reference data (Figure 3.):

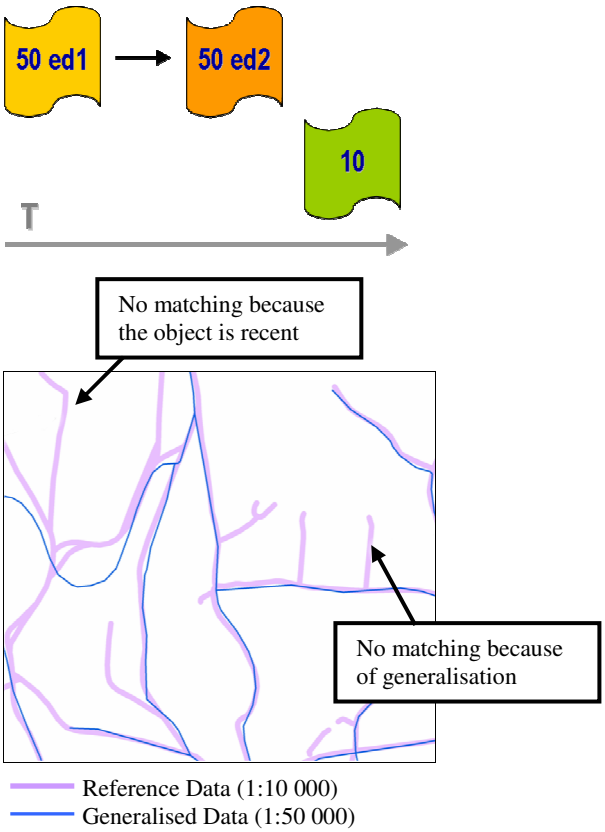

Figure 3. Situation 1

In this first situation (concerning only a few 1:50 000 sheets), we will take advantage of the linking process to identify the update differences and synchronise the data simultaneously.

2. Reference data is older than generalised data (Figure 4.):

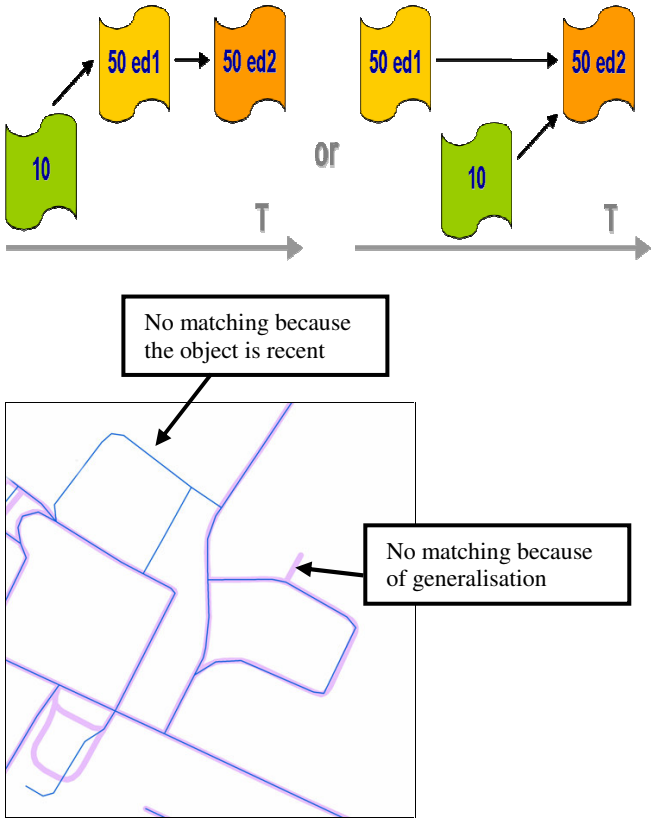

Figure 4. Situation 2

In this second situation, which is the usual case, the updated generalised data will certainly be used to indicate where the reference data should in priority be updated but, for resolution reasons, updating the "old" reference data with the newer generalised data will only be possible for some specific features or attributes.

One solution to the synchronisation problem could be to postpone the matching (and of course the updates propagation process) after the updating of the reference data. Then we would be back in the situation presented above. But this would cause a longer delay between the 1:10 000 updated data production and the derived map production.

To save time, another solution would be to perform the matching process now without synchronising the data and then process the updates propagation as soon as the reference data will have been updated. In that case, however, only part of the reference data updates will have to be propagated.
The difficulty will be to identify this part. Indeed, if we only store the successful matching relations, there will be no direct way to check automatically, for example, that a new object

already exists in the generalised data as no matching was found when linking the data. In order to facilitate the future updates propagation, an instance will be created in the relation table even for the unmatched features to record the reason for which the matching did not occur.
The records stored in the relation table will have to be characterised to allow the distinction between the different kinds of matching failures illustrated above (update and generalisation). We could for example, define an attribute "Relation type" with the following values:

- Matching
- Matching but an attribute was modified
- Matching but the geometry was modified
- Without link because of deletion
- Without link because of creation
- Without link because of generalisation

Even if the fact that a feature is relevant or not for the generalised data can always be checked on the fly according to a set of rules, it would probably be quicker to retrieve the information directly from the relation table.

When updates propagation will be performed, the set of reference data updates will have to be checked against the last five types of feature relation types to detect if the update has already been applied to the generalised feature, in order to avoid errors such as duplications.

4. CREATING THE LINK BETWEEN CORRESPONDING FEATURES IN THE DIFFERENT DATABASES

The automatic matching of features can rely on several algorithms using semantic information (e.g. attribute items), geometric information (e.g. shape) and topologic information (e.g. association) of the geographic feature. The algorithms first allow to select the candidate features to link and subsequently will allow to validate (approve or reject) the created associations. The complexity consists in the choice and order of the algorithms.

The generic process according to (Badard and Lemarié, 2002) and (Devogele, 1997) is the following:

- The geographic datasets are enriched by adding attribute items or characteristics on the shape of the features (e.g. lines coming in/out of a node).
- The candidates are selected by an area around the geometry of the source feature (e.g. buffer) and/or by measuring a distance between features (Hausdorff, Fréchet, etc.).
- An association is created between the selected features and the source feature. This is done by measures based on several tools and their right parameters. Priority is given to features with the same semantic information.
- Measures or tools will detect irrelevant features and delete their association with the source feature.
- Sometimes the candidate features will be extended when the association appears to be unreliable. In this case, the measures are triggered again with new candidate features.
- The features are grouped by their association and features appearing in different groups are detected.
- The validity of each association is checked (manually, but guided by the process).

At IGN Belgium, we developed an ArcGis application in Visual Basic to link the 1:50 000 road network with the corresponding features in the 1:10 000 data. The described application relies only on a geometric analysis using buffers around the geometry of the 1:50 000 feature. First, with several parameters, we analyse the 1:10 000 candidate features completely inside the buffer. If no match is found, 1:10 000 features partially inside the buffer are analysed. For each matching pair, the object id's are stored in a relation table.

Figure 5. next page gives an overview of the script.

The following parameters are used:

- *(1) Buffer maximum*: The maximum size that can be reached during the increment of the buffer (+5m) around the 1:50 000 feature. The goal of this parameter is to limit the research area in order to have only the relevant candidate features and to save processing time.
- *(2) Curve parameter:* The ratio between the shortest distance between the two extremities of the line and the length of the line. This parameter is equal to 1 for a straight line, and 0 for a line for which the two extremities meet. In all other cases, the value of the parameter will vary between 0 and 1. This parameter allows to skip the parameter *maximum angle* when the curvature of the line is too elevated.
- *(3) Maximum angle*: The maximum angle between the 1:10 000 candidate feature and the part of the 1:50 000 feature inside the buffer around the 1:10 000 candidate feature. The goal of this parameter is to reject 1:10 000 features perpendicular or transversal to the 1:50 000 feature.
- *(4) MinAbsLength:* The minimum total length of a 1:10 000 feature in the buffer around the 1:50 000 feature.
- *(5) MinPropLength*: The ratio between the length of the 1:10 000 feature contained in the buffer around the 1:50 000 feature and the total length of the 1:10 000 feature. This parameter is equal to 1 when the 1:10 000 feature is completely inside the buffer and 0 when the 1:10 000 feature is completely outside the buffer. In all other cases, the value of the parameter will vary between 0 and 1. When the proportion of the 1:10 000 feature inside the buffer is lower than the minimum value allowed for this parameter, the candidate feature is skipped .
- *(6) RapDistanceMin:* The ratio between the sum of the lengths of the valid 1:10 000 candidate features inside the buffer around the 1:50 000 feature and the length of the 1:50 000 feature. So long as this ratio is lower than the minimum value allowed for this parameter, the size of the buffer will increase until the size of *Buffer maximum* is reached.

A test was done on 1397 1:50 000 road segments. 95% of them were correctly associated. These results are very promising, but some refinements still have to be done (Darras, 2004; Determe, 2005):

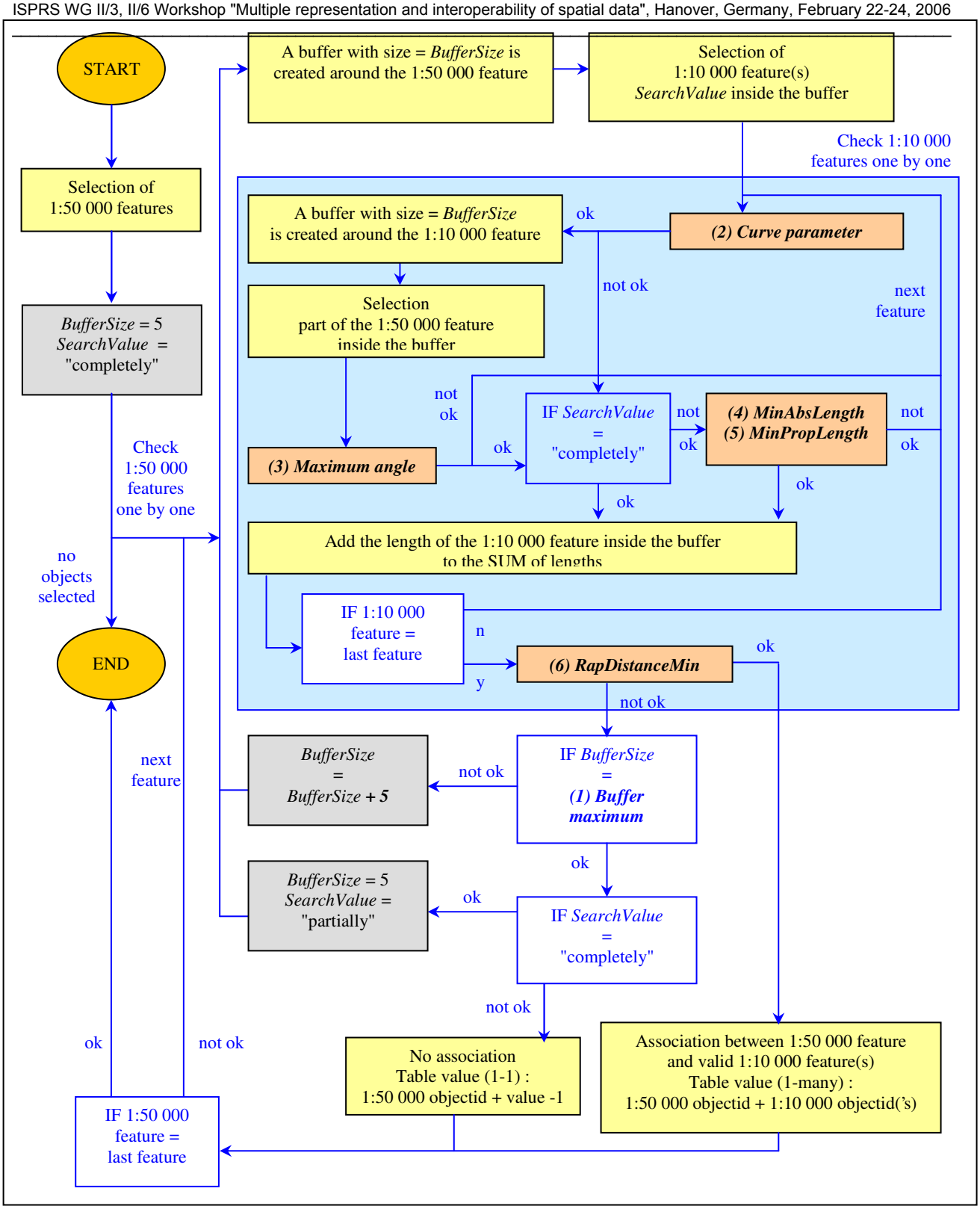

Figure 5. ArcGis application to link the road network in the 1:50 000 database with the corresponding features in the 1:10 000 database

- Topology should be taken into account when linking corresponding features. The validation test (Determe, 2005) showed that many of the remaining errors could be avoided if the relative position to the nodes was considered.
- At this moment, there are no 1:10 000 candidate features found when the 1:50 000 feature is displaced over a distance larger then *Buffer maximum*. This is rather uncommon, but the following solution could be used: if the parameter *RapDistanceMin* is between 0.75 and 0.9 and the BufferSize is 15m, the *Buffer maximum* should increase (up to 25 or 30m) until *RapDistanceMin* is reached. Other constraints should probably be added to avoid inadequate associations.
- Roundabouts and dead-ends with a loop at the end in the 1:10 000 data should be identified and handled in a specific way.
- When *RapDistanceMin* is exceeding 1.25, it is likely that irrelevant 1:10 000 features are linked. Therefore, we need to define priorities to order the 1:10 000 candidate features when calculating the sum of the lengths.
- The limit values of the parameters were fixed empirically. The statistics (max, min, mean, …) calculated for each parameters in the set of correctly matched features identified in the test could help to fix the critical values in a more efficient way (Determe, 2005).
- The process stops when *RapDistanceMin* is fulfilled. In some cases we can continue the process by increasing the *BufferSize* or by using the *SearchValue* = "partially". Trying the remaining candidate features could improve the matching.
- We also intend to test JCS Conflation Suite against our home software for the automatic matching (Stigmar, 2004).

The link between corresponding features will be stored in a relation table where the cardinality problem will be handled in the way shown in figure 6.

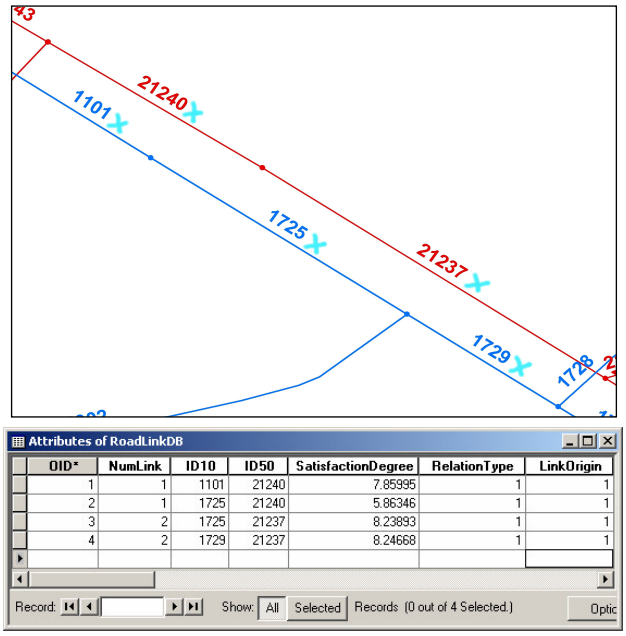

Figure 6. The relation table

In the relation table, besides the id's of the corresponding features, an attribute storing the way that the link was created (automatic matching non-checked, automatic matching checked, manual matching or generalisation) was added (Anders and Bobrich, 2004).
At this moment, nothing can assure the reliability of the associations between the 1:50 000 and 1:10 000 features. We have no other choice than the manual check. However, a global indicator, based on the satisfaction degree of some of the parameters calculated during each linking process, could help us to retrieve the suspicious features. This indicator is stored in the relation table to guide the manual checks. The way to calculate this indicator is not final. It still has to be improved (Determe, 2005).

5. CONCLUSION

In a few years, IGN Belgium will have to produce up-to-date generalised maps from a large–scale reference dataset as automatically as possible.

The ideal way to produce those maps would be the full automatic generalisation of the updated reference data, but this option is not yet realistic. The process could be automated gradually, but spending hours of interactive generalisation in a semi-automatic process while the generalised data will already be existing (in a previous edition) would not be very efficient.
Starting a generalisation process all over again only makes sense in our case if it requires less interactive work than the updates propagation solution. It could be the case for some features, like the very changing ordinary buildings that must be typified.

Part of the updates propagation process can be automated quite rapidly. Moreover, this option has other advantages like the data quality control.
The integration of the datasets at 1:10 000 and 1:50 000 scales is now being studied. Some conflicts have been identified, among which the update date difference, which complicates the feature linking process and consequently the first updates propagation.
A VBA script for the road matching has been written and tested. It still has to be improved. Among other things, topology should be taken into account, the script should be more flexible and the limit values of the parameters should be determined using the statistics calculated in the set of correctly matched features.
We also have to improve the global indicator reflecting the quality of the matching, in order to reduce the number of manual checks without any risk. Then the script will have to be adapted to the other types of features.

The automatic generalisation is investigated in parallel and is certainly the long-term preferred solution. Even in the updates propagation solution, it will be useful because the updates have to be generalised. It could also be necessary for the production of other derived products.

6. REFERENCES

AGENT, 2000. Project AGENT, ESPRIT/LTR/24939. http://agent.ign.fr

Anders K.-H. and Bobrich J., 2004. MRDB approach for automatic incremental update, ICA Workshop on Generalisation and Multiple representation – 20-21 August 2004, Leicester.

Badard D. and Lemarié C., 2002. Associer des données : l'appariement in *Généralisation et représentation multiple*. Hermès Science Publications, Paris.

Darras B., 2004. Appariement de bases de données topographiques, Rapport de stage, D.E.S. en cartographie et télédétection, Année académique 2003-2004.

Determe K., 2005. Evaluation du résultat de l'appariement automatique, Rapport de stage, D.E.S. en cartographie et télédétection, Année académique 2004-2005.

Devogele T., 1997. Processus d'intégration et d'appariement de bases de données géographiques – Application à une base de données routières multi-échelles, Thèse de doctorat en méthodes informatiques, Université de Versailles, décembre 1997.

Ruas A., 2002. Pourquoi associer les représentations des données géographiques in *Généralisation et représentation multiple*. Hermès Science Publications, Paris.

Stigmar H., 2004. Merging Route Data and Cartographic Data, ICA Workshop on Generalisation and Multiple representation – 20-21 August 2004, Leicester.