Managing the National AR5 Dataset in a Loosely Coupled Distributed Heterogeneous Database Environment

Knut Bjørkelo and Håvard Tveite

The Norwegian Forest and Landscape Institute
Knut.Bjorkelo@skogoglandskap.no
The Norwegian University of Life Sciences
havard.tveite@umb.no

AR5 is the most detailed land cover dataset in Norway, with a minimum mapping unit 500 square meters. The AR5 dataset classifies all land of Norway into several classes with the main purpose of representing current land use and production potential for forest and agricultural crops.
The main AR5 classification is on surface type, and there are supplementary classifications on forest production potential, forest cover type and soil condition.
AR5 is a discrete coverage dataset consisting of polygonal units that is a result of a human classification based on observations of a complex set of biological and environmental factors that vary more or less continuously in space.
The Norwegian municipalities (400+) are responsible for the daily maintenance of the AR5 dataset, according to its specification. The national dataset is obtained by collecting the master databases from the municipalities periodically. This involves harmonisation, with respect to topology and classification, of the (local) master databases to achieve national coverage.
We investigate different solutions to our main challenge: How can we efficiently maintain a seamless national AR5 dataset from hundreds of autonomous, heterogeneous database systems distributed throughout Norway?

1 AR5 – contents and management

AR5 is the most detailed land cover dataset in Norway [12]. The SOSI 4.0 [1] data model (UML) of the AR5 dataset is shown in figure 1.

AR5 is a discrete coverage [6] dataset that classifies the non-urban land areas of Norway into several classes with the purpose of representing current land use and production potential for forest and agricultural crops. The dataset consists of approximately 10 million polygons with a total of 1 billion points in their borders, and covers the complete area of Norway. The minimum mapping unit is 200 square meters for agricultural land, and up to 0.5 hectares in less productive areas.

The following classes are defined in AR5 for area types ("arealtype" in the UML model): *Built up areas and infrastructure* (built-up, infrastructure), *agriculture* (crop-land, surface grown crop-land, grazing-land), *forest*, *open land*, *glacier*, *water*

(freshwater, ocean) and *not mapped*. The classes in parenthesis are specialisation classes.

There are separate classifications for forest production potential (7 classes), forest cover type (6 classes) and soil condition (7 classes). The other classifications listed in the UML class diagram in figure 1 are not in use.

There are 2 types of boundaries for the AR5 units, regular boundaries between different area types, and virtual borders imposed to reduce the size of polygons. In addition, the regular boundaries are tagged to distinguish borders "inherited" from other datasets.

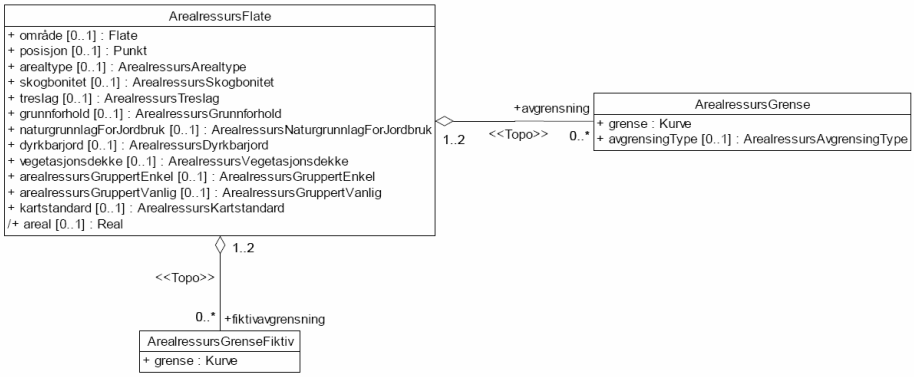

Fig. 1. UML model for the AR5 dataset (from SOSI 4.0 [1])

AR5 is a discrete coverage dataset [6] consisting of polygonal units that result from a human classification based on observations of a complex set of biological and environmental factors that may vary more or less continuously in space. The polygonal units with their classifications and their boundaries are very much a result of human interpretation. Since the biological and environmental factors that are observed vary continuously over space, the unit boundaries are not crisp. An indication of the degree of fuzziness of an AR5 boundary is available as 4 visibility classes of the boundary attribute SYNBARHET. Both the classification of the units and the delineation of the units have elements of fuzziness [7]. So, if two persons classify an area, the resulting polygonal units and their classifications will most probably not be identical. Boundaries for roads and water polygons are, to a large extent, imported (copied) from other datasets which are maintained separately by other responsible parties.

Updates to the AR5 dataset will in some cases cover several square kilometers and due to some interactive work it will result in very long transactions.

1.1 Managing AR5 – responsibilities

The Norwegian municipalities (400+) are responsible for the daily maintenance of the AR5 dataset, according to the AR5 specifications. Many municipalities are cooperating on geographical data management, and have formed groups for that purpose, so the number of AR5 databases is significantly less than 400.

The current procedure for maintaining the national AR5 dataset is based on the assumption that the master database is managed in a distributed manner by the municipalities.

The Norwegian Forest and Landscape Institute [5] has the coordinating national responsibility for the AR5 dataset. The (seamless) national dataset is currently obtained by collecting the master databases from the municipalities periodically. This involves harmonisation of the (local) master databases to achieve nation-wide coverage. In the harmonisation process, special attention must be paid to the areas in the vicinity of municipality boundaries. The harmonised national AR5 database is sent back to the municipalities by clipping it using the municipality borders, applying the official administrative borders provided by the national mapping authority for clipping.

1.2 Municipal boundaries

The official municipality borders are maintained by the Norwegian mapping authority [4]. The municipality borders are not static. Structural changes occur where two or several municipalities are combined into a new, larger municipality. Municipality borders are also, for various reasons, adjusted from time to time.

The municipality boundaries are currently also represented in the AR5 dataset (as unit boundaries). The AR5 "version" of these will be slightly different from the official version, so the clipping approach will introduce artificial polygons (slivers), way below the minimum mapping unit, resulting in geometrical and topological problems. The changes may be at millimeter level due to coordinate transformation, meters due to various kinds of revisions, or larger. Sliver polygons down to the minimum possible triangle in the coordinate realm may occur, leaving no place for a representation point and provoking software failure.

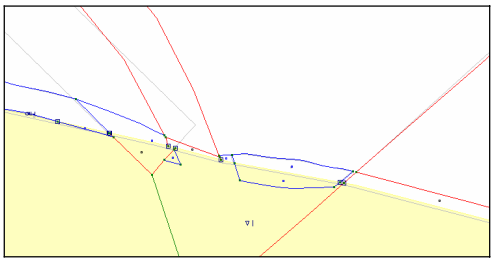

Fig. X. Parallel lines and slivers introduced along municipality border.

The national AR5 dataset is supposed to be seamless, so AR5 units should, in principle, not be split across administrative boundaries. The mapping of AR5 should focus on the landscape, disregarding any administrative borders, and neighbour municipalities should cooperate when AR5 is revised in the border zone. A bonus of the seamless approach is that the sliver problem at the municipality borders is eliminated, avoiding unnecessary work and software problems.

2 Database system options for the AR5 dataset

There are several alternatives for organising a database system [8] which may be applicable to the AR5 data management.

- The traditional centralised database system where the database management system (DBMS) and the database is hosted by one organisation at one location. All clients work directly on this database system using a computer network and standardised interfaces or clients. A centralised database system must support concurrency control for multiple users in the presence of long update transactions. The transaction lengths for AR5 could be up to several weeks! Locking must be supported at several levels of granularity and area locks are very useful.
  This solution is used for the Norwegian "Matrikkel" (Cadaster) system, but at the moment this is not a viable solution for AR5.
- A homogeneous distributed database management system (DDBMS) consisting of a set of nodes, often at different locations, connected by a computer network. All the nodes run the same database management system software (or support a powerful standardised interface for distributed database management). Data are distributed among the nodes using horizontal or vertical fragmentation [2].
  In the case of AR5, horizontal fragmentation (database tuples are distributed among the nodes) based on spatial location would have to be the distribution method. Concurrency control and transaction support must be provided, as for a centralised system, but the distributed nature of the system makes transaction support more challenging.
  The Norwegian municipalities do not have access to software that supports a homogeneous DDBMS platform, and the AR5 project does not have the powers to dictate the municipalities choice of database management solutions.
- A federated DDBMS consisting of independent DBMSs with a loose coupling. The nodes can organise the data the way they like, but there should be an overall interface to the complete database.
- Independent local DBMSs, where an integrated database is only generated periodically, and a lot of harmonisation is required for data that are shared between two (or several) different nodes.
  At the moment this is the situation for the AR5 database. Harmonisation is required at the municipality boundaries, and much of this work has to be done manually and requires a lot of resources.

A standard interface for vector geographical data, like WFS transactional, can be used to access (for retrieval and update) a distributed geographical database system as a whole, or could be used in a federated geographical database system for access to the individual participating database systems by the federation layer.

3 AR5 coordination / integration

The main challenge of AR5 management is to maintain a seamless national AR5 dataset that use data from a large set (hundreds) of autonomous heterogeneous database systems distributed throughout the country.

Several widely used applications require that a seamless national coverage, with the most recent updates, is easily available. The central distribution database may now be one year behind on updates, even more for some regions. As the local update frequency of AR5 data improves further the lack of efficient integration methods becomes critical.

The main sources of complexity for the AR5 database are the (non-static) municipality borders and the non-crisp nature of the AR5 "features". At the municipality borders the responsibility and authority change from one municipality to another.

Further complexity is imposed by the current transition from the old national coordinate system (NGO48) which has a problem with deformations, to a new coordinate system based on the European datum (Euref89) and UTM projections.

Schema harmonisation is, in principle, assured through compliance with the SOSI 4.0 standard. However, the different GIS tools and DBMSs in use will to some extent modify representation of attributes and geometries to their native concepts, and this may complicate harmonisation.

In this presentation we leave the coordinate system and schema challenges in the background, and focus on the municipality borders.

3.1 Obtaining a seamless AR5

If two municipalities are allowed to update the AR5 dataset along their common boundary totally independently, the municipality boundary will become a seam (see figure 2). Classifications will change at the municipality boundary and AR5 unit boundaries will not be harmonised.

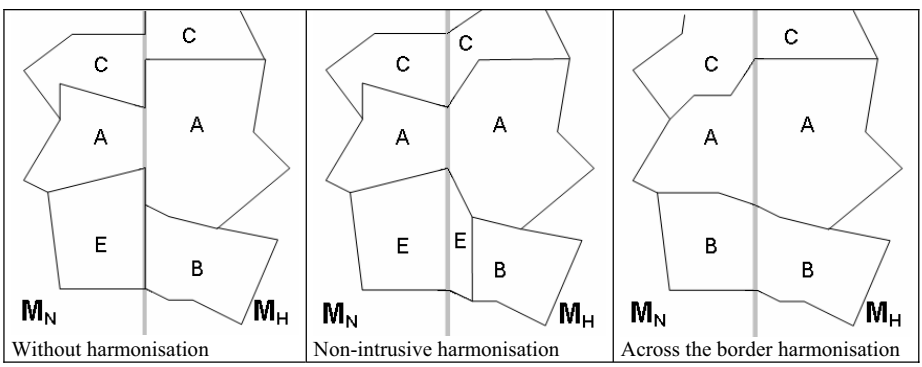

Fig. 2. Harmonisation at municipality boundaries

To avoid the seam at the municipality boundaries, a municipality will have to access the AR5 databases of its neighbouring municipalities. The AR5 classification can then be harmonised along the boundaries, obtaining continuous AR5 units over the municipality boundaries. The AR5 creation / update process will then have to be constrained by the existing AR5 classification in the neighbouring municipalities in order to minimise the effects on the topology of the involved databases.

The simplest solution is only to consider AR5 units in neighbouring municipalities that have a boundary that is partly made up of the municipality boundary, and to consider their boundaries and classification as fixed. This approach has the drawback that it will not allow the classification to be modified at the municipality boundaries. This means that either the units will not generally be continuous over the municipality boundaries (figure 2, left), or that "artificial" narrow "units" are introduced along the municipality boundaries with a classification that corresponds to the "old" classification of the other municipality, or that erroneous classifications are exported from the neighbouring municipality into units at the municipality boundary. The second approach is illustrated in figure 2 (middle), with the "home" municipality (M_H) to the right of the thick grey line (the municipality border), and the AR5 classes in small capital letter (A, B, C, E).

A more complex, but certainly better, solution is to include a transition zone into the neighbouring municipalities, so that the new classification will not have to be constrained by faulty older classifications at the municipality boundary. In this case, AR5 units in neighbouring municipalities that are direct neighbours to the municipality can have their boundaries and classifications changed, while units that are not direct neighbours should be considered as fixed. One variant of this approach (not disturbing the boundaries of the second order neighbours) is illustrated in figure 2 (rightmost), with the "home" municipality (M_H) to the right of the thick grey line (the municipality border).

General approaches to topological continuity handling in geographical multi-database systems encompass correspondence establishment, semantic conflict resolution, geometric conflict resolution, boundary alignment, topological continuity and more [11]. In the AR5 case, correspondence establishment will be complicated by the fuzzy nature of the AR5 units, but fortunately, the area where correspondence establishment is necessary is small (units that are affected by the municipality borders).

3.2 Concurrency control and harmonisation

There are several alternatives for concurrency control in a DDBMS. Locking (two-phase locking - 2PL), time stamp ordering and optimistic methods are important classes [9]. For long transactions, as in the AR5 case, concurrency can be improved by using some kind of versioning [10], combined with optimistic methods and conflict resolution.

The future architecture of the AR5 will probably be a combination of a centralised DBMS solution where some municipalities will choose a direct read and update access to the centralised DBMS, while other municipalities will choose a detached solution with a local master AR5 DBMS and periodical harmonisation with the nation-wide DBMS. The frequency and time of import could vary from municipality to municipality.

Currently the central database has a simple mechanism that mark, at the object level, data that are copies from a detached original DBMS.

For the sake of inter-municipality harmonisation, it would be advantageous if the national AR5 database could support multiple versions of the database. To support

the harmonisation process, the different versions should be identified by metadata: a) the municipality responsible for the update; b) time of update.

Direct access to the centralised DBMS

For the direct access solution, traditional concurrency methods could be applied, taking into consideration long update transactions and the harmonisation and authority challenges at the municipality boundaries. Geographical area locking would be preferable, but updates (at least those along the municipality boundaries) should be managed as new versions (not overwriting the current version) to allow later harmonisation with the detached DBMSs at the municipality boundaries. Using versions, locking can also be minimised and concurrency improved.

General purpose database management systems do not provide area locking, but object level locking and dataset / table level locking will normally be available. Area type locks could be supported by some kind of spatial predicate locking in a geographically enabled database management system, or must be supported indirectly using object level locking by middleware or the application itself. The WFS 1.1 [3] specification only supports object level locking.

Detached DBMSs with periodic harmonisation

The direct application of traditional real-time concurrency control methods is probably not realistic for the detached DBMSs. The main objective of AR5 concurrency control should be to allow the municipalities to update their part of AR5 without too much hassle, while being able to maintain a consistent nation-wide dataset by periodically integrating and harmonising the different local DBMSs into the central AR5 DBMSs. The AR5 update method should as far as possible allow the harmonisation process to be performed automatically.

AR5 concurrency control could be split into two levels – the local (municipality) level and the national level.

At the local level, when someone is going to update the AR5 dataset, a lock will have to be set on the geographical area where the update is going to take place, in order to avoid database inconsistencies due to parallel local updates. Locks can be acquired on individual objects, or on a geographic area. For a coverage type dataset, area locks are more attractive than object locks due to the arbitrariness of the units and the unit boundaries.

At the national level, the AR5 can be viewed as a collection of datasets that are updated individually. The partitioning (horisontal fragmentation) of the AR5 is done based on municipality boundaries, but to support the harmonisation process, a local dataset should be able to extend across these boundaries into the neighbouring municipalities.

Locking is not an option at the national level, due to the lack of mechanisms for synchronisation between the local DBMSs. Optimistic methods with the use of versions in the conflict resolution process could be an option. Each update received from a local DBMS could then be treated as a separate version, tagged with the municipality ID and the time of update. The harmonisation process would then take into account the last consistent version of the nation-wide AR5 dataset and all the new versions received from the municipalities, and perform the harmonisation of the AR5

units at the municipality boundaries according to the ideas outlined earlier in the paper.

The extent of the local datasets' geographical coverage could be constrained to support one of the harmonisation methods discussed above.

Municipality boundaries

To allow the handling of the effects of changes to municipality borders and municipality structure during harmonisation, it is important that all relevant versions of the municipality borders, with their period of validity are available to the centralised national AR5 DBMS, and perhaps also to the local DBMS.

Even if no administrative boundaries are represented in the dataset, versioning would be a good approach for the municipalities cooperation at their borders.

Harmonisation

In cases where two or more municipalities have updated the same area, harmonisation is necessary. Automatic harmonisation might be possible in cases where classifications can be unified using conflict resolution, and the new AR5 boundaries can be interpolated from the available versions. In some cases, automatic conflict resolution will not be possible, and manual handling would be needed. We will not go into the details of conflict resolution.

When municipality boundaries change, the area of authority changes too. During conflict resolution, the time of the boundary will have to be taken into account. If there are authority conflicts due to changes in boundaries, they will probably have to be resolved manually.

3.3 Consequences for users of the national AR5 dataset

Users of the national AR5 dataset should have easy access to the best version of the seamless coverage. The complexity of distributed databases, versioning and conflict resolution should not affect applications using the national AR5 dataset. The alternatives for organising the database system pose different challenges.

The centralised database system gives one single point of access, and can easily be fitted with interfaces for data retrieval to satisfy clients. If the central database contains several versions or unresolved conflicts, a server side application layer could encapsulate the complexity, hiding irrelevant details from external developers / users.

Likewise, a homogeneous distributed database could handle data collection through its nodes. To avoid situations where incomplete or erroneous data are sent to clients, an advanced topologically aware server side application is required.

A federated database will further extend the challenges for the server side application. The risk of being unable to deliver correct data to clients must be handled.

Independent local databases may make it practically impossible to supply correct data from one central server.

As stated before, independent local databases is the current situation for the AR5 original dataset. Instead of putting efforts into developing an application that access

the local databases and tries automatic data harmonisation, the Norwegian Forest and Landscape Institut has chosen to establish a central database with a mix of original data and copy data. This way clients are guaranteed access to correct and seamless data. However, the data may not show the latest updates that have occured in the detached original (primary) databases.

4 Conclusions

A discrete coverage dataset with conceptually non-crisp elements resulting from several classifications based on field type variables poses challenges to data management in the presence of long update transactions. A further complication with the AR5 dataset is that it will be managed as a combination of a centralised database system and a distributed set of autonomous municipal level databases, resulting in only periodical updates of the centralised database system. Updates to the AR5 will normally cover many square kilometers and will result in very long transactions. We discussed methods that can limit the efforts needed in the harmonisation process, based on a focus on the municipality boundaries and the topological properties of the units that are affected by the boundaries.

In order to make automatic harmonisation possible, it is suggested that all updates that concern AR5 units that are affected by municipality boundaries are handled using versions. This will allow all relevant information to be available in the harmonisation process. Within the municipality boundaries, update authority is not split, and traditional concurrency methods can be used.

Until efficient tools for automatic harmonisation of independent local databases are available (at a reasonable cost), The Norwegian Forest and Landscape Institute has decided to establish one central database with a mix of original data and copy data. They encourage the local authorities responsible for update of AR5 to store their original data in the central database. In order to further reduce the technical challenges with updates from local originals, cooperation in the border zones between local databases is encouraged.

References

1. *SOSI 4.0*. The Norwegian Mapping Authority.
   URL: http://www.statkart.no/sosi/welcome.htm
2. Ceri, S. and Pelegatti, G. *Distributed Databases – Principles & Systems*. McGraw-Hill, 1984.
3. OGC. Web Feature Service (WFS). Open Geospatial Consortium Inc.
   URL: http://www.opengeospatial.org/standards/wfs
4. The Norwegian Mapping Authority. URL: http://www.statkart.no
5. The Norwegian Forest and Landscape Institute (Skog og Landskap).
   URL: http://www.skogoglandskap.no
6. ISO. *ISO 19123 Geographic information - Schema for coverage geometry and functions*, International Organisation for Standardisation, 2005.

7. Burrough, P.A. *Natural Objects with Indeterminate Boundaries*. In *Geographic Objects with Indeterminate Boundaries* (editors Burrough, P.A. and Frank, A.U.), volume 2 of GISDATA series. Taylor & Francis, 1996, pp 3-28.
8. Elmasri, R. and Navathe, S.B. *Fundamentals of Database Systems*, Benjamin/Cummmings, 1989.
9. Bernstein, P.A., Hadzilacos, V. and Goodman, N. *Concurrency Control and Recovery in Database Systems*. Reading, MA, Addison Wesley, 1987.
10. Korth, H.F. and Speegle, G.D. *Formal Aspects of Concurrency Control in Long-Duration Transaction Systems Using the NT/PV Model*. ACM Transactions on Database Systems, Vol. 19, No. 3, 1994, pp 492-535.
11. Laurini, R. *Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability*. International Journal of Geographical Information Science, Vol. 12, No. 4, January 1998, pp 373-402.
12. Bjørdal, I. and Bjørkelo, K. *AR5 klassifikasjonssystem. Klassifikasjon av arealressurser.* Handbook from Skog og landskap, 01/2006.