The New and Improved Base Framework For National Atlas Data

Rupert Brooks
Canada Centre for Remote Sensing, Geomatics Canada
Brooks@NRCan.gc.ca

Abstract

Reliable framework data is essential for the practical use of geospatial data. It is one of the objectives of Geo-Connections (formerly the Canadian GeoSpatial Data Infrastructure or CGDI) to provide reliable framework datasets for Canada. The National Atlas of Canada is addressing this need by developing integrated and reliable framework datasets based on the Vector Map Level 0 (VMAP0) product.

The VMAP0 data product is a global digital dataset produced and maintained by the United States National Imagery and Mapping Agency (NIMA) at a nominal scale of 1 : 1,000,000. The National Atlas, in wanting to choose an international global digital map base, has adapted the VMAP0 for use as a base from which other, smaller map scales can be created and to which thematic layers can be added.

The 1:1M data had to be processed and quality controlled to allow it to be used. Since the National Atlas represents data at a national level, a variety of derived scales are being produced from the VMAP0 data. This is being done using automatic model generalisation techniques, as developed by Richardson [1996]. To prepare the data for these techniques, however, existing topology had to be corrected, and even more rigorous topology had to be defined.

Having overcome these shortcomings, the National Atlas of Canada VMAP0 data for Canada is of the highest quality. In addition, geostatistical atomic units of area such as Census Subdivisions have been fit to an enhanced VMAP0 base at the scale of 1 : 1,000,000, which should make the dataset even more useful. Throughout these processes, rigorous quality control has been applied to ensure the best possible result.

This paper discusses both the techniques used to achieve these results, and the quality control processes used to yield products of the highest quality.

Introduction

In order to better relate its datasets to global initatives, and to better meet the needs of base data users at Atlas scales, the National Atlas of Canada is reconstructing all its bases at a nominal scale of 1:1,000,000. (The word nominal is used because the concept of scale has changed somewhat with the advent of digital maps.) To date, the hydrology and boundary layers have received the most work. These layers have been derived primarily from the VMAP (Vector Map) Level 0 revision 4 (hereafter, VMAP0) data. It is anticipated that other layers will come from a variety of sources.

The VMAP0 data was first released as the DCW (Digital Chart of the World) dataset. The DCW has been widely used and has proved to be very useful because it was the first global multitheme dataset at such a scale in the public domain. Despite its success, the DCW dataset is widely known to have had some deficiencies. The revision 4 data for North America was much better in quality, but still had to be improved before it would be suitable for use by the National Atlas of Canada.

Two kinds of improvements were performed on the dataset. First, improvements were performed where the data was occasionally in error, out of date, or incomplete. In particular, some parts of the James Bay power project and the Nunavut territory were not properly represented. International marine boundaries were not part of the dataset, and had to be added. Certain rivers which were connected in reality were not

connected in the database. These types of improvements acted to update and correct the VMAP0 data, but did not appreciably change its specification. A second type of improvement was an improvement in structuring. The 1:1M base, when it is complete, will need to be generalised for use at smaller scales. The National Atlas of Canada has neither the time nor the personnel to do a manual generalisation of the entire country, so automated generalisation techniques will be used. These techniques require a highly topologically structured dataset, and considerable effort has been expended on topological structuring. These enhancements exceed the original specifications of the VMAP data.

The shift in bases is occurring simultaneously with a philosophical shift in purpose. Historically, the Atlas has produced cartographic products intended for viewing on paper, at a fixed scale, with the cartographer retaining ultimate control over the presentation of the data. This process is being supplanted by the production of datasets allowing interactive mapping. This new paradigm allows much more power for presentation, but at the same time places many more demands on the cartographer. It requires rethinking the concepts of scale, generalisation and quality control.

In no way should this new attitude be construed as meaning that cartography is no longer required. In fact, it is needed more than ever before. The Atlas is providing (and always has provided) ***information*** products rather than ***data*** products. The difference is that information products are expected to convey their meaning directly. An Atlas map should not have to be corrected or adjusted by its user, and should not have artifacts of the data in it. Since a user may – is expected to - combine the individual themes of data in unanticipated ways this places an obligation to have individual datasets be compatible to a high degree of accuracy. This requirement is further reinforced by the fact that many Atlas users are not geomatics professionals. In order to provide meaningful information to the user, the data must be of the highest quality.

Scale

The concept of scale has changed greatly with the advent of interactive mapping. In the past, paper maps would be designed for intense study. Therefore detail was held to be accurate to the limits of the medium, but errors smaller than what could be seen on the paper could be ignored. On a computer screen, two factors combine to produce a new way of interacting with a map. Firstly, the resolution is much lower, so it is physically impossible to pack as much information into the same space. Secondly, the computer user will tend to zoom in on areas of interest, rather than meticulously studying them at the original scale of presentation. Therefore, there is a change in the meaning of the scale parameter, and an increased and changed requirement for generalisation. Rather than building a series of separate, independent products at different scales, the National Atlas of Canada is now interested in building a single integrated geospatial database that supports representation at a variety of scales.

Multiscalar Datasets

A user will in general have more interest in some areas than in others. Canada is a country with a wide variation in the distribution of its people. Vast tracts of land are essentially unpopulated, while well over a third of the population is concentrated in or around the four largest cities. For any socially related theme, the feature density available and required in urban areas far exceeds that available for the remainder of the country. On a paper map, this sort of information would be handled by inset maps. On an online map representation, this must be handled by zooming. In the paper case the inset map could be, and usually was, quite separate from the less detailed main map. In the online case, there is a requirement for very detailed information in urban areas, integrated seamlessly with the smaller scale remainder of the dataset.

The National Atlas of Canada is experimenting with a multiscalar approach where the base layers in these areas of great interest will have much greater detail. The effective scale in these areas will be 1 : 250,000 or larger but they will be seamlessly integrated with the surrounding data at 1 : 1,000,000. These areas of high detail will serve the function that inset maps served in a paper cartography world. They will be spatially compatible with the 1 : 1,000,000 data that surrounds them, and their feature density will be gradually tapered towards the edge of the area rather than abruptly changed. This seamless integration will allow them to be used in online mapping.

This approach has been used in the past. In fact, the census subdivision (CSD) boundary dataset, as distributed by Statistics Canada, has long exhibited this sort of structure. It was the challenge of creating a base consistent with these multiscalar datasets, that prompted the effort to develop a consistent set of multiscalar base data.

Generalisation:

The National Atlas of Canada is responsible for strategic mapping at a wide range of scales, and because of these wide variations in level of detail, generalisation is a essential capability. Generalisation is commonly broken down into two main types, model generalisation and cartographic generalisation. Both types are necessary to produce a good map representation [João, 1998].

Model generalisation involves the selection of the right features for display at a particular scale. This varies in difficulty depending on the type of data involved. For populated places, for instance, the population of a place is a good guide to whether or not it gets displayed at a particular scale. For a feature class with strong internal relationships between its features, such as hydrology, or geology, these issues become quite complex. For instance, river systems are generalised by removing the fingertip channels first, and working in towards a central main stream, which will be kept at all but the highest levels of generalisation [João, 1998].

Cartographic generalisation involves the adjustment of the presentation of the selected features to make the display intelligible at the desired scale. While this aspect of generalisation is subtler than the outright selection of features, it is also very important to the generation of meaningful maps. For paper products, linework must be kept to certain degree of smoothness to avoid ink bleeds. For the screen, while the pixelation does automatically generalise features to a certain extent, because of the limited resolution, exaggeration and displacement become more important. Also, the size in bytes of the dataset is an important factor in the response time of an interactive online system [João, 1998].

The National Atlas of Canada does not intend to manually generalise a base dataset for the entire country. Automated generalisation will be used to produce base layers at smaller scales than 1:1,000,000. Richardson [1996] has developed automated generalisation routines for hydrology and transportation networks that can be used to perform the required model generalisation. Cartographic generalisation routines that preserve topological information have been developed (eg. Saalfeld, [1998]), but to date no specific procedure has been chosen to use on this data.

The automated model generalisation techniques work through feature selection based on importance computed by an analysis of the network topology. In effect, they explicitly model the relationships a human cartographer perceives intuitively. Clearly, for these to work correctly, it is vital to have accurate network topology. While the original VMAP was topologically structured, the accuracy of this structuring was not up to the exacting standards of accuracy required. Furthermore, the VMAP specifications did not include any requirement for accurate directionality of river systems. This is, however, needed for the generalisation algorithms to work. Therefore, a considerable amount of the processing of the data has involved detecting and correcting topological problems, and more will be done in the future.

Quality control

While the National Atlas of Canada has always taken a great deal of pride in the accuracy of its products, the new paradigm in online mapping has placed even greater requirements on quality control. As users now have the capability to use the data in unexpected ways, subtle errors or inconsistencies that could have been controlled in a paper environment can now become serious irritations. The construction of the base layers has been subjected to rigorous quality control procedures, and careful records have been kept. Each change to the dataset can be tracked and justified [Brooks et al., 1999].

Definitions of quality

At small scales, such as 1:1,000,000 and below, a user is less interested in positional precision than in accuracy of meaning. For instance, it is much more meaningful for a typical atlas user to know which islands belong to Canada, and which belong to the United States along the boundary between the two countries, than to know the boundary coordinates to sub-meter precision. On the other hand, it is the policy of the National Atlas of Canada to adhere to the highest available positional accuracy. When integrating new information not only must the immediate problems of data integration be dealt with, but also the indirect problems due to relationships affecting the updated features.

There are a number of relationships both in and between different geographic objects in the database which need to be correctly maintained. For instance, a river may also be an international boundary, which may also be the boundary of a census division, and of an ecological zone. It is no longer adequate to merely have a visual fit. Exact and explicit matching is needed if a dataset is to be suitable for analytical use.

This relationship is especially important because with the available software mapping tools, the user can view and query the data in far more sophisticated ways than ever before. Mistakes will subtly transmit their problems into these views and queries. They will either be detected, to the embarrassment of the author, or, worse still, they will mislead the user.

There are stronger relationships between certain types of geographic objects than between others. These stronger relationships allow identification of several families of datasets. Within a family, there are many strong relationships between the different object, while the relationships between objects in different families are loose. For instance, political boundaries, census subdivisions, and electoral districts form a strongly related family. Census subdivisions and electoral districts share many of their boundaries, and all provincial and national boundaries also form the boundary of some census subdivision, and some electoral district. Rather than treating each of these as a separate layer of information, they are all being integrated into one dataset, where the relationships between boundaries will be implicit.

Finally, when assessing the quality of a dataset, it is important to take its lineage into consideration. Wherever possible, the final product is based on data obtained as close as possible to its source. In cases of confusion, care was taken to verify the dataset against another with a different lineage. For instance, the VMAP0 data and the existing Atlas 1 : 2,000,000 hydrographic base were derived from the same source data. It was no surprise that a high degree of agreement was found between them. A stricter verification was to compare the VMAP0 data to the National Topographic Data Base (NTDB) map sheets which were separately compiled, sometimes from entirely different aerial photography. This type of verification was done for any change or displacement of a feature over a distance greater than two kilometres.

Data processing

The procedural and philosophical shifts discussed above required that we update, correct and enhance the VMAP0 datasets before using them as base data. To date, the hydrology and boundaries datasets have been significantly improved.

The hydrology data was chosen as the first layer to process for two reasons. The VMAP0 hydrology data makes up over half of the total VMAP0 data (almost half a million arc features) and by handling it first it was possible to better ascertain the quality of the data. It is also one of the most accurate layers in that dataset, for two reasons. Firstly, as the original purpose of the source dataset was for aeronautic navigation, rivers took on a high level of importance being used for visual reference by pilots. Secondly, hydrology is one of the slowest base layers to change, which meant that the late 1980's vintage of the data did not present a problem.

The boundaries layer in the VMAP0 dataset also includes the coastline. In the National Atlas of Canada model of hydrology, however, the coastline is essential. For this reason, the boundary layer was processed concurrently with the hydrology layer. Duplicate copies of the coastline are preserved in both the new

hydrology and boundary datasets. The new boundary dataset was based mainly on the boundary information from VMAP0. It was also adjusted by the addition of more accurate or more detailed information from other sources. In certain cases, more accurate information actually required the displacement of the VMAP hydrology in order to preserve the relationships between features.

Both layers of the data went through the following basic steps. First the format was converted from the Vector Product Format (VPF) in which they were distributed. Data conversion almost inevitably causes some data loss due to incompatibility in the data models. Great pains were taken to minimize this data loss Even using the VPF to Arc/Info conversion tools which are available in Arc/Info, the program for making this conversions ran to well over 1000 lines of code [Brooks et al. 1999].

Following the import, the feature coding and other logical consistency checks were performed. In this phase, errors such as rivers running through the middle of lakes were detected and corrected. Tile boundaries were identified for later removal, and a duplicate coastline was placed in each of the boundary and hydrology datasets.

The two datasets then followed somewhat different processing paths. The hydrology data, because of its greater complexity, and because of the demanding requirements of the automated generalisation routines had to be checked meticulously for topological errors.

The main type of topological error checked for was the connectivity of river systems to the coastline. These errors were detected by making the assumptions that rivers generally drain to the ocean, and that river systems are generally acyclic graphs. Neither of these assumptions is entirely true in practice, but because they are true in the vast majority of cases the false alarms generated by these checking procedures were small. Also, these are the same assumptions used by the generalisation algorithms so deviations from them needed to be detected for special treatment in any case. Analytical tools were used to detect both the case where a small gap was keeping a river system from connecting as it should, and the much rarer case where two river systems were connected together that should not be.

As a support to these checks, the existence of large and medium scale water diversions in Canada was researched. These diversions tended to disrupt the acyclic topology of river systems in which they occurred. Therefore they were identified and integrated into the overall data model. Similarly, internal drainage systems, defined as river systems that do not drain to the ocean, have been identified and their sinks determined. In the beginning, almost 35% of the arcs in the dataset were not attached to any sink, after careful checking, that figure has been reduced to about 12%. This reduction was achieved by adding less than 3000 short connecting arcs. Bearing in mind that a certain proportion of the arcs in the dataset are intended to be isolated lakes and streams, that figure is considered good.

The boundary data needed less topological processing than the hydrology data. The data needed to be updated with the addition of the new Canadian territory, Nunavut. As well, the National Atlas of Canada had access to a very accurate international boundary and international marine boundary information, and this was used to enhance the dataset. In certain cases where the international boundary lay along a river, the addition of more accurate information also affected the hydrology layer. Care was taken to be certain that the information represented in both datasets is precisely identical.

The finest level of administrative subdivision which existed in the VMAP0 data set was the provincial boundaries. The National Atlas of Canada frequently uses census subdivision polygons for displaying social thematic information collected on these divisions. Census subdivisions are also considered part of the same data family as the provincial boundaries. Accordingly, these were integrated into the boundary dataset.

When processing the data, great care was taken to make sure that the data was genuinely enhanced, and not corrupted. When arcs were added to correct connectivity problems in all but the rarest of cases they were checked against another base map. Each such case was recorded, is traceable and the justification for its existence can be determined. At each stage of processing, the number of arcs was carefully tracked.

Furthermore, automatic generalisation routines that could move or delete features were avoided in favor of manual checking. While tedious, this process has payed off by producing a dataset of the highest quality.

These datasets have been greatly enhanced, but further progress is still necessary. In order to prepare the hydrology for automatic generalisation, some further topological processing will be necessary. In particular, the directionality of each river in the dataset will have to be determined. More administrative and political divisions will be added to the boundary layer. Further datasets will have to be processed in order to create a set of minimal framework data layers. At a minimum, these datasets will include the recommended [Evangelatos, 1999] elevation, toponymy, transportation and imagery layers. The National Atlas of Canada is currently exploring a number of options, but at this point it appears unlikely that the data for these layers will come exclusively from VMAP0.

Summary

The National Atlas of Canada is building a set of bases at a nominal scale of 1:1M. These bases are intended to be both suitable and compatible with mapping of the nation or parts of it, but also suitable for integration with global mapping projects. To date, a hydrology and boundary dataset have been produced which are primarily based on VMAP0 data. This data has been updated, corrected and enhanced by the addition of new information, careful inspection, and analytical techniques.

Overall, the dataset proved to be of high quality from the outset. The original dataset contained on the order of half a million arcs; this number decreased a bit after processing. Of the various types of errors investigated, all could be said to occur for less than 1% of the features in the dataset. This starting error rate has been greatly reduced by the application of quality control procedures.

In the future, the National Atlas of Canada will continue to refine and construct base map layers at the 1:1M scale. These datasets will be constructed in such a way that the topological relationships between features, and the relationships between datasets will be explicitly maintained. This will be necessary if the datasets are to be automatically generalised in the future.

References

Brooks, R. Evans, R. Huang. J. 1999. Processing of the VMAP Hydrology layer. Internal report of the GeoAccess Division.

Evangelatos, T. 1999. IACG / CCOG National Coordination Meeting on CGDI. Summary notes from CGDI Framework data conference (February, 1999). Available from http://cgdi.gc.ca.

João, E. 1998. *Causes and Consequences of Map Generalisation.* Monograph. Francis and Taylor.

NIMA, 1995. *Military Specification: Vector Smart Map Level 0.* Military Standard number MIL-V-89039. National Imagery and Mapping Agency, United States of America. http://www.nima.mil.

Richardson, D. E., 1996. Automatic Processes in Database Building and Subsequent Automatic Abstractions. In *Temporal, Spatial, and Semantic Data Integration for Application in Remote Sensing and Geographic Information Systems.* Dianne E. Richardson, Ed. *Cartographica*. Monograph, Fall 1996.

Saalfeld, A. 1998. Topologically Consistent Line Simplification With the Douglas-Peucker Algorithm. Preprint paper from the Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University. Columbus OH.