INTEGRATION AND HARMONIZATION OF CONTRIBUTIONS TO A EUROPEAN DATASET

Andreas Illert, Ingo Wilski
Institut für Angewandte Geodäsie (IfAG)
Richard Strauss Allee 11
D-60598 Frankfurt am Main, Germany

Abstract

The joint usage of datasets from different sources requires the harmonization of the data with regard to the data model, the data structure and the georeference. The difficulties with this task are pointed out by the example of a European dataset of administrative boundaries created from national contributions.

1 Introduction

With the number of suppliers for geo-related data increasing continuously, the potential user has the chance to collect information from a large variety of sources. On the other hand the setting of tasks to GIS is getting so complex that technical data from several disciplines have to be combined and evaluated together. At the same time the number of GIS projects covering more than one national territory is steadily increasing. All these trends aim at the common exploitation of different data, thus requiring their harmonization and integration to a common model. Typical of the input data in a process of harmonization is their heterogeneity with regard to geometric quality and/or semantic modeling .

First practical applications of computer-assisted harmonization concentrated on the adaption of inhomogenous geometry with compensating distribution of residual gaps [e.g. 1,2]. In the meantime, approaches get into the focus of interest which do not consider only the possible discrepancies in the geometry but also the semantic differences due to a deviating model conception. Such differences in conceptual modeling do not only occur with datasets covering different themes. Because of a lack of national and international standards, even datasets that cover a common theme like topography do hardly refer to compatible models if they originate from different producers.

If the input data to a harmonization procedure differs with regard to the resolution, the concepts of generalization have to be considered. Furthermore, the problems related to data integration are closely connected with the aspects of data quality [3]. For example, knowledge about the quality of the data is essential with the harmonization of those objects which are common to several datasets. With regard to automatization the knowledge must be formalized. At present the description and exploitation of semantic knowledge is still a matter of research. Therefore, only a few parts within the procedures for the linkage of heterogeneous datasets will be subject to automatization for the time being. The status and the applicational possibilities of computer-assisted data harmonization shall be pictured in the following by an example.

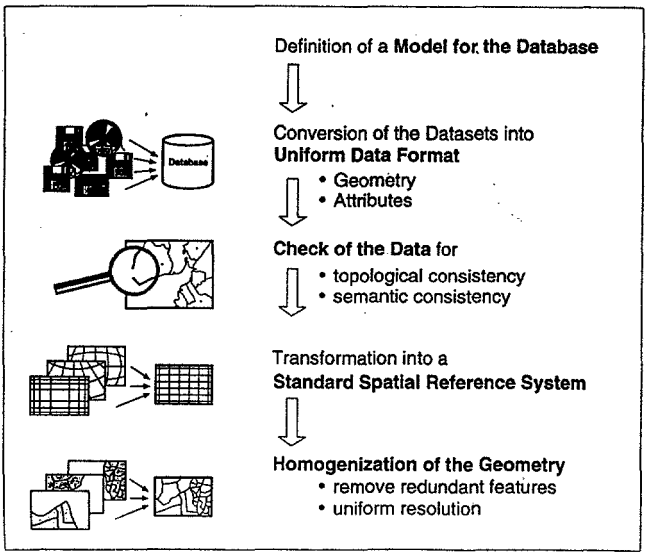

Figure 1: Working steps for the harmonization of the administrative boundaries

2 An example : the harmonization of datasets of European administrative boundaries

The Institute for Applied Geodesy (IfAG) at Frankfurt am Main / Germany is treating, within the scope of its activities and as Service Centre within the MEGRIN project, the harmonization of datasets of European administrative boundaries. MEGRIN (Multipurpose European Ground Related Information Network) is a project of C.E.R.C.O. (Comité Européen des Responsables de la Cartographie Officielle) which aims at the creation of a technical and organisational framework to simplify the provision of geographic information [4]. Users in need of transnational data are to be supplied according to uniform standards through this network. Besides the creation of an information system with metadata, this concept also allows for the development of seamless datasets over Europe. A pilot project with regard to future MEGRIN products is the Seamless Administrative Boundaries of Europe dataset (SABE). The realization of this dataset is dominated by the principle of not performing a reacquisition of the basic data but to found on the digital resources of the respective official national mapping agencies. Thus, much importance is attached to the aspects of data integration and harmonization.

The national datasets made available to the MEGRIN Service Centre cover nearly the whole of Europe, with the exception of some states in East and South-East Europe. By January 1995 the collection consisted of more than 20 national contributions, with the territory of Germany split further into 13 separate datasets supplied by the individual state survey administrations. The geometric resolution of the SABE dataset is defined corresponding to the map scale 1 : 50 000 . The dataset is to represent the administrative structure down to the lowest administrative unit having an elected parliament and a budget of its own (commune level).

After the collection of the data the Service Centre has a mosaic of heterogeneous datasets at its disposal which in part differ considerably from each other :

- The data was digitized from maps with scales ranging from 1 : 5 000 to 1 : 1 000 000.
- The spectrum of the data modeling extends from simple plotfiles to a sophisticated database.
- The resolution of the data does not always reach down to the commune level. Sometimes it can neither be clearly seen from the available information which level within the national structure constitutes the lowest one possessing an elected parliament and a budget of its own.
- The geometry of the data refers in most cases to the national geodetic datum and to a specific map projection.
- The data is delivered in different exchange formats.

Processing at the Service Centre aims at producing a homogeneous and consistent dataset on the basis of a standardized model with uniform spatial reference. The relevant activities are performed at IfAG through the MEGRIN Service Centre. Harmonization proceeds according to the following steps :

1. Definition of a model for the database of the administrative boundaries
2. Conversion of the datasets into a uniform data format
3. Check of the data for topological and semantic consistency
4. Transformation of the geometry into a standard spatial reference system
5. Harmonization of the geometry at common boundaries

These single steps shall be examined more closely in the following.

3 Definition of a model for the database

The joint utilization of the various datasets requires employment of a common model. This model must be suited for expressing the different administrative structures of the European countries. It has to represent relations such as the aggregation of administrative units into structures of higher order. Special cases, as for example exclaves, condominiums or boundaries lacking a clearly defined geometry should pose no problems. The search for a common denominator is complicated by the fact that the administrative structures on which the respective datasets are based differ strongly. The data models correspond, as a rule, to the national conditions and cannot be transferred without difficulties to other ones.

In order to meet the demands a model is designed which conforms to CERCO's European Territorial Data Base (ETDB) specifications. The model founds on the surfaces of the lowest-level administrative units from which the national territory can be composed as a mosaic. Each area is described geometrically by the linear segments of the bounding polygon and a centroid. The higher-level administrative levels are described indirectly by the areas of the assigned lowest-level units. They refer to a lowest-level administrative unit as a seat of the administration. The model described has been implemented at IfAG on the ARC/INFO system.

4 Format conversion

The datasets available to the MEGRIN Service Centre were supplied in almost a dozen different exchange formats. Further processing requires conversion into a common data format. In the process the available information must be transformed into a form which corresponds to the model defined before.

As target system ARC/INFO with the exchange formats EXPORT and GENERATE is selected. While the data converters contained in the ARC/INFO package satisfy most demands with respect to the geometry, the transformation of the attributes and the relations turned out to be by far more difficult. Any attempt at automatic conversion failed when semantic information is coded incompatibly to the model or, at worst, by graphic aids such as arrows of association or text elements in vector form. Names, designations and numerical codes must therefore be extracted to a large extent manually.

5 Check of the data

The individual datasets must be checked for consistency and completeness. This requirement applies to the geometric and topological quality as well as to the semantic quality of the data. A thorough check would call for a comparison with reality. However, the Service Centre has hardly any independent sources at its disposal by means of which such a comparison would be possible. In this case a check can only be performed in a rudimentary manner. However, deficiencies which may lead to conflicts with the conditions of the model must in any case be removed, as e.g. the requirement for precisely one centroid per area.

In practise, the check of the data is oriented towards special error situations. These error situations have to be recognized already at the conception stage. Then test methods may be developed to detect the errors in the single datasets.

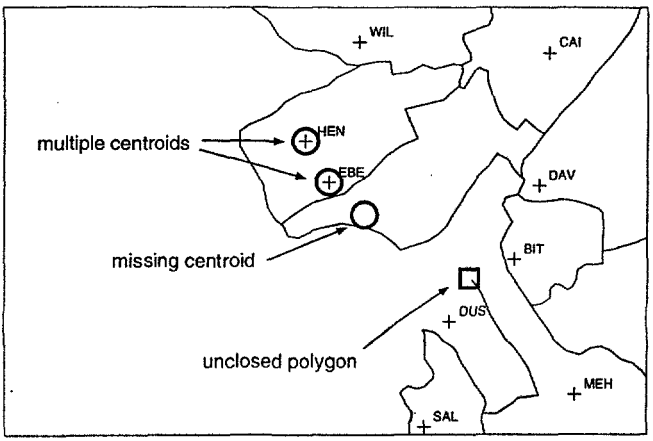

Figure 2 : Topological and semantic deficiencies

For the recognition of errors in the topology and the geometric consistency, GIS systems like the ARC/INFO provide a number of suitable tools. By means of such tools the geometry can be checked, among others, for unclosed polygons or redundant lines. For instance, such deficiencies occur in case that coast or shore lines have not been considered (figure 2).

Contrary to the topology, only a small part of the semantics can be verified by automatic techniques. The predominant portion of the checking procedure for semantic plausibility and completeness is performed manually by operators at displays or on paper plots. The detection of errors depends on the knowledge and the experience of the operator. Apart from that, a large part of the discrepancies discovered in the datasets requires consultation with the data suppliers. In many cases it becomes apparent that there is no error, but that the set of special and particular cases within the respective administrative structure is enriched by a further aspect.

The process of data checking is not restricted to the time cited here in the flow of data harmonization. On the contrary, the control of the data is performed during and after each single step. On the one hand, each step of processing may reveal further errors. On the other hand, the data editing may affect the datasets in a way that previous checking procedures have to be applied again.

6 Transformation into a standard reference system

The datasets refer with very few exceptions to national reference systems. They do not only differ with regard to the geodetic datum but are also subjected to specific map projections. For common transnational use the data have to be transformed into a standard spatial reference system.

The target system chosen are geographic coordinates on the GRS80 ellipsoid referring to datum WGS84 or ETRF89. Transformation of the datasets to this standard reference system is carried out in four steps :

1. The plane cartographic coordinates have to be transformed into geographic coordinates referred to the national ellipsoid. For this purpose knowledge about the type of the map projection used, its specific parameters and the national ellipsoid is required.

2. The national geographic coordinates are to be transformed into national rectangular coordinates. For this purpose knowledge about the parameters of the ellipsoid used is required. As result are obtained Cartesian coordinates referred to the national datum.

3. The national rectangular coordinates are to be transformed into a standard reference system (WGS84 or ETRF89 in our case). For this purpose knowledge about the parameters of the displacement vector, the rotations and the scale difference is required. These parameters are, as a rule, not published by the responsible survey administrations, but in most cases at least approximated values are obtainable which serve cartographic purposes.

4. Finally, the international rectangular coordinates are to be transformed into geographic coordinates referred to the standard ellipsoid GRS80.

From the standard reference system the harmonized dataset can later be retransformed into any cartographic projection by inversion of the transformations.

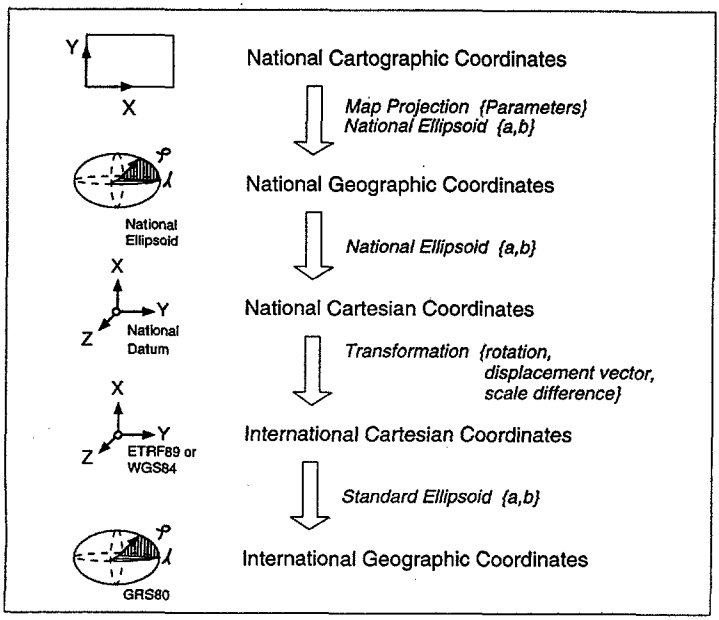

Figure 3 : Transformation into a standard reference system

7 Harmonization of the geometry

If the single datasets are finally available in a uniform data model and with uniform spatial reference, they have to be linked together such that a consistent non-redundant dataset is created, and that the appearance of the integrated geometry is as homogeneous as possible.

As the national datasets are strictly limited to the respective national territory, redundancies in the adaption process occur only at the common borders. These have, as a rule, already been attributed as national boundaries and can therefore be easily detected. In each case the geometry is taken from the dataset with the larger scale of the digitizing original, assuming that this is the dataset of higher accuracy. The topology of the other dataset is adapted to this geometry. A local adjustment is not performed. If the two border lines differ strongly from each other, the cause of the discrepancy must be determined in accordance with the data suppliers. Figure 4 illustrates such a situation in the border area of two states A and B. At first sight the reason for the deviation seems to be the stronger degree of generalization of the national border of dataset A. This contradicts with the fact that A has used the more accurate digitizing original, which is manifested in the details of the internal boundaries. The datasets possibly refer to boundary definitions at different times. This example proves the necessity of thourough investigation and illustrates the difficulties occuring with the search and analysis of errors.

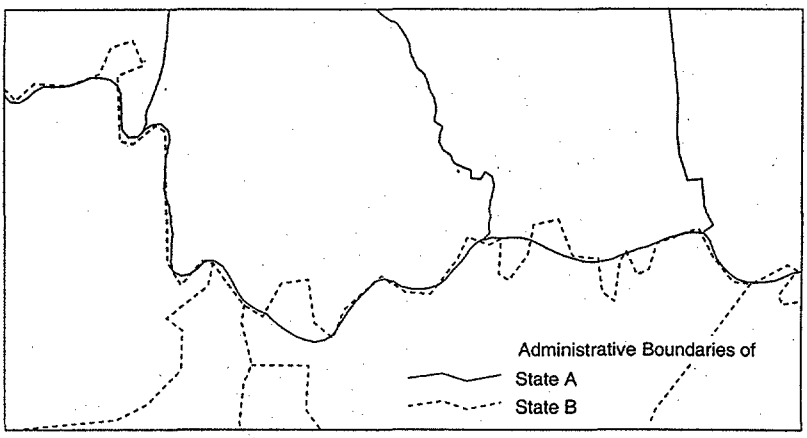

Figure 4 : Inconsistent geometry at the common boundary of two datasets

The appearance of the national datasets shows considerable differences which are due to the enormous variation at the scale of source maps. Harmonization of the appearance therefore includes generalization of the geometry insofar that it has been acquired from maps of higher resolution than the target scale. To reach this goal at minimal costs a line simplification algorithm will be applied without regard for complex correlations. In our case the Douglas-Peucker algorithm [5] is used. After simplification the geometry must again be checked for topological and semantic consistency. More complex operations of cartographic generalization, as e.g. enlarging of small but important structures, are not aimed at in order to keep the interactive portion of work as low as possible.

8 Automatization

With regard to the large amount of data the harmonization procedure is to be automatized as far as possible. However, it turned out that the software tools of modern GIS systems allow only for parts of this task. As an example, the ARC/INFO system provides suitable tools for the cartographic projections, the conversion of geometry between standard data formats, and the recognition of errors in the topology. Automatic tools lack with the model transition of attributes and the recognition of errors in the semantics. In some cases one may have recourse to the programming of special check routines. These routines must be tailored to each individual case and to the respective data model.

As a rule, the automatic tools offered by GIS systems for the harmonization of data are restricted to the geometry. As the processing of attributes and relations is closely related to the specific application, general solutions are not available at present. It is on the user to realize his specific conversion from an input dataset to a uniform model.

9 Conclusion

As compared to other possible applications of data integration the project of harmonization of European administrative boundaries seems quite simple. Redundant geometry occurs only on the national borders, no sophisticated adjustment is performed with the geometry, and complex cartographic ge-

neralization is referred to only marginally. Performance of the project nevertheless entails considerable efforts. The limits of the automatization become clearly evident even with such a clear case as are the administrative boundaries.

The provision of meta information for the harmonization of the heterogeneous datasets is very helpful. Any of the datasets hardly could have been processed succesfully without additional knowledge about the type of the digitization original, the procedure of data acquisition, and the national administrative structures. Such meta data is not only important to the Service Centre as data editing agency, but is also of great importance to the potential user. Just in case that a dataset is derived from different sources the description of the particular sources and the operational process constitute a decisive basis for judging the data quality.

Automatization of the harmonization procedure is not only appropriate with regard to an acceleration of the procedure and the economy of manpower, but ensures also the reproducibility of the process. On the contrary, the results of human interaction can be reproduced only to a limited extent. This may turn out to be embarrassing at a later update. However, at the present status of development the use of automatic tools is possible only to a limited extent. Especially the processing of the semantic information still requires a high degree of interaction. Here, automization can only be achieved by means of specific programming for the specific transition from the national dataset to the standard model. A more flexible software requires formalization of model descriptions and semantic information.

Harmonization at the scale described here can hardly be expected of a potential user. Considering the necessary efforts, he will no doubt decide himself against the heterogeneous sources and prefer a complete reacquisition under his own responsibility. In order to prevent this the suppliers of digital information must either include harmonization as a service in their offer, or keep from the beginning strictly to uniform standards for models, data format and interface.

10 References

[1] Jessip, M.B., 1991. Systematic horizontal adjustment of positional error in digital line graphs. Technical papers, 1991 ACSM-ASPRS Annual Convention, Vol. 2, Baltimore, USA,

[2] Morgenstern, D., Prell, K.-M., Riemer,H.-G., 1988. Digitalisierung, Aufbereitung und Verbesserung inhomogener Katasterkarten. Allgemeine Vermessungs-Nachrichten, No. 8-9/88, pp. 314-324

[3] Guptill, S.C., 1993. Describing spatial data quality. Proceedings 16. International Cartographic Conference, Cologne, Germany, pp. 552-558

[4] Salgé, F., 1995. Standardization in the Field of Geographic Information : The European efforts. this proceedings.

[5] Mc Master, R.B., 1987. Automated Line Generalisation. Cartographica, Vol. 24, No. 2/87, pp. 74-111