Identification and adjustment of corresponding objects in data sets of different origin

Monika Sester, Guido von Gösseln, Birgit Kieler
Institute of Cartography and Geoinformatics, Leibniz University of Hannover
{monika.sester, guido.vongoesseln, birgit.kieler}@ikg.uni-hannover.de

Abstract
The paper tackles the problem of data integration. It presents first results of an approach for the instance-based semantic integration. Subsequently, the geometric integration is presented using a method for homogenization that allows for a partial integration taking different geometric deviations into account. Results are shown with the example of the integration and adaptation of data of different origin from topographic and navigation data sets.

Introduction and overview

The availability of spatial data sets, especially also over the internet via web services, allows the combination of data of different origin. However, geometric overlay of data often reveals semantic and geometric differences – even when the same or similar objects in reality are concerned. In order to be able to have an integrated analysis, the data sets have to be fused or homogenized. The fusion process needs some basic prerequisites, which will be tackled in the paper: first of all, semantically corresponding objects have to be known. Then, instances of these objects have to be identified which can be overlaid. Finally, the boundaries have to be mutually adjusted in order to obtain one synchronized object geometry. Based on a set of corresponding objects which serve as "ground control objects" or anchor objects, also the rest of the data set, i.e. the non-anchor objects can be transformed.

Having such a system realizes the vision that data sets can be selected and loaded for a certain application, followed by an automatic mutual adjustment, so that a coherent geometry is determined as a basis for follow-up analysis and visualization processes. In the ideal case this is done on-the-fly. The adaptation can be considered as a kind of synchronization process. For this synchronization rules have to be defined that specify how objects have to be altered in the fusion process. Thus these rules control e.g. whether one object is representing a "master geometry" to which the objects of the other data set are adjusted, or alternatively, whether an intermediate geometry between both data sets has to be determined.

The paper is organized as follows. In the next section the background of the research is sketched as well as references to existing work are given. Then, the first results of the semantic integration is described. Subsequently, our method for the geometric adaptation is shown. A summary and an outlook conclude the paper.

Background and state of the art

Interoperability and especially data integration faces different types of problems (Bishr, 1997): it has to take structural, semantic and geometric differences in the data sets into account. Structural interoperability can be achieved using standardized data formats (e.g. ISO, OGC). The harder problem is semantic interoperability, aiming at integrating data from different domains and communities. Kuhn (2003) coined the metaphor of a "semantic reference system" that should be set

up – comparable to a geodetic reference system, where transformations between semantic spaces and projections to sub-spaces are defined.

The basic questions to be solved in order to achieve the vision of interoperability are:

- Which objects do correspond ?
- How should corresponding objects be geometrically adjusted ?

Geometric matching or harmonization has been tackled since some decades. One of the first approaches was by Saalfeld (1988), who also coined the notion of "conflation" for this process of matching and geometric data fusion. There are many approaches for geometric matching that mainly differ in the type of transformation that is assumed between the two data sets, ranging from identical reference system, i.e. no transformation, to geometric differences in terms of translation, scale and rotation. In the latter case, matching procedures that take also relations into account have proven to be the most flexible (Walter & Fritsch, 2000). For the geometric adaptation, rubber sheeting algorithms have been proposed where in the simple case a vector field is calculated from some given object correspondences (Doytsher, 2000). Other approaches allow for the integration of geometric constraints within the objects, which have to be preserved in the transformation process (e.g. Hettwer, 2000).

Before, however, geometric objects can be matched first of comparable objects or object classes have to be identified. To this end, similar object schemata have to be known. Kokla (2006) presents guidelines for geographic schema or ontology integration. One option to perform this identification of semantic correspondences is to do it manually using a careful inspection of given object catalogues or ontologies. An integration of different ontologies can be achieved by mapping terms of individual ontologies to a common shared ontology. It can also be modelled using the concept of multiple representation. Balley et al. (2004) describe an approach which aims at a loose coupling of different schemas by specifying correspondence relations between object classes of the different ontologies. Vangenot (2004) creates a common integrated schema and uses the notion of "stamps" to differentiate the different appearances of the objects in the different domains. Stoimenov & Djordjevic-Kajan (2002) describe a framework for semantic interoperability defining possible relation concepts like equivalence, subsumption, overlap or difference. In the EU-project GiMoDig, which aimed at an on-the-fly harmonization of topographic data sets in Europe, a common scheme has been identified, together with rules that describe how to transform the object classes of the individual countries into the common scheme (Sarjakoski et al., 2002).

Such a manual process is adequate, as long as dedicated integration processes for a limited number of data sets are looked for. If we, however, aim at an integration of arbitrary data sets that can be loaded in the internet, then this manual semantic translation is no longer be feasible. In order to automate this process, a so called instance-based or extensional determination of schema transformation rules can be used (Kokla 2006, Duckham & Worboys, 2005). The underlying idea of this approach is as follows: if two objects have an identical name and / or geometrically coincide, then they probably also have something in common on the semantic level. In the general case, one cannot assume, that the names of the objects in different data sets are the same – except objects with a unique given name like names of cities or roads. Thus, we only use geometric relations to infer a semantic relation. This approach has been used by Volz (2006) to link linear data.

The possible discrepancies between the data sets depend on the application they have been acquired for, the level of detail (scale), and on the acquisition itself. Different modelling schemes may, e.g., lead to different aggregation levels of the objects and thus to different cardinalities between the object instances in the different data sets. E.g. a set of lakes in one data set can be modelled as an aggregated lake in the other. If the data sets stem from different scales, more severe differences between them can occur and have to be modelled, namely larger geometric discrepancies and also topological differences. In this paper we start from the assumption that the data sets to be integrated are of similar scale. This implies, that also the boundaries between the objects are similar.

Description of used data sets

For this work two topographic data sets, GDF and ATKIS data, were used: whereas GDF data (Geographic Data Files) was specially developed for purposes of vehicle navigation and is captured for most areas in Western Europe, ATKIS (Authoritative topographic cartographic information system) data provides a basic set of topographic objects and is available for the whole area of Germany in four different scales. In our work, we used the most detailed scale, the so-called Base-DLM of scale 1:25.000. GDF is of the same scale. The test data set lies within the urban area of Hanover and has a size of 25 km^2. GDF as well as ATKIS model objects with all geometric types: points, lines and polygons. For this investigation only polygon-objects from both data sets were used. Table 1 lists the object groups and the quantities of objects in the respective groups. The obviously different number of objects in both data sets indicate different modelling techniques, namely that the ATKIS data are modelled in much greater detail than the GDF data, or, vice versa, that GDF contains many aggregated objects.

data set	object groups		total
GDF	wa (water area)	water element	10
	lu (landuse area)	ft: 7170 park, garden	1
		ft: 9715 industrial area	6
	lc (landcover area)	woodland, moor and sand	3
	a8 (administrative area)	municipality	2
ATKIS	2000 (urban area)	e.g. industrial area	754
	4000 (vegetation)	e.g. grassland, arable land, forest, moor	226
	5000 (water area)	e.g. stream, river, brook, pond	66
	7000 (administrative area)	different order (e.g. municipality, city, town)	2

Table 1: Used object groups and quantities from GDF and ATKIS data sets.

Identification of similar semantic objects

Our goal is to identify corresponding semantic object groups like lakes, woodland or roads in the different data sets. In the general case we cannot assume that expert knowledge about the precise meaning of the terminology used by the organizations which capture and modell the data sets is always available. Therefore, this should be accomplished by the analysis of geometric and topological characteristics of the objects. The quality of the results of the determination of objects with similar semantics depends on the similarity of the data sets. In our case, we restricted ourselves to objects of the same geometric type, namely polygon-objects.

As described above, we apply a bottom-up approach to accomplish an identification of similar geometrical elements. A simple geometric overlay with the data sets of a same geographical extent is used. This intersection often returns more than one matching objects: on the one hand this is due to the fact that there are always geometric discrepancies on the object boundaries. On the other hand, the objects are not modelled in a tessellation, i.e. that at one position there is always only one unique object class. This is especially true for administrative objects that often encompass larger areas. To restrict the possible matching and improve the results the additional criterion *size* is introduced.

Particularly the ratio *(R)* of the intersection area *(I)* and the object *size (O)* as illustrated by figure 2 and equation 1 is analysed to get further information about possible matching-partners.

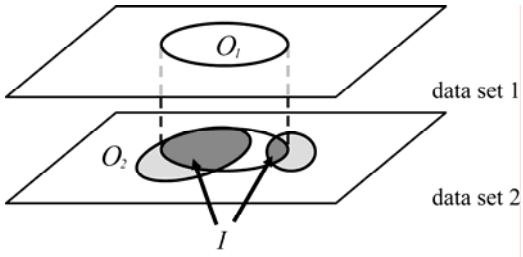

Figure 1: Ratio between intersection area *(I)* and object size *(O)*.

$$R_i = \frac{I \cdot 100\,\%}{O_i} \qquad \text{with } i = 1,2 \qquad (1)$$

Taking slight geometric differences into account, we consider objects to match, when the ratio is 80% or better. In the case of a 1:1-relationship, this 80%-ratio has to hold for both matching directions, i.e. for i=1 and 2. The search for 1:1 relations in our example data set returned only four objects within the 80% ratio. As described above, obviously GDF contains more aggregated objects Therefore the investigation was extended from 1:1 relations to 1:n relations. In that case the 80%-ratio between the areas has to be met *at least* in one direction of the investigation.

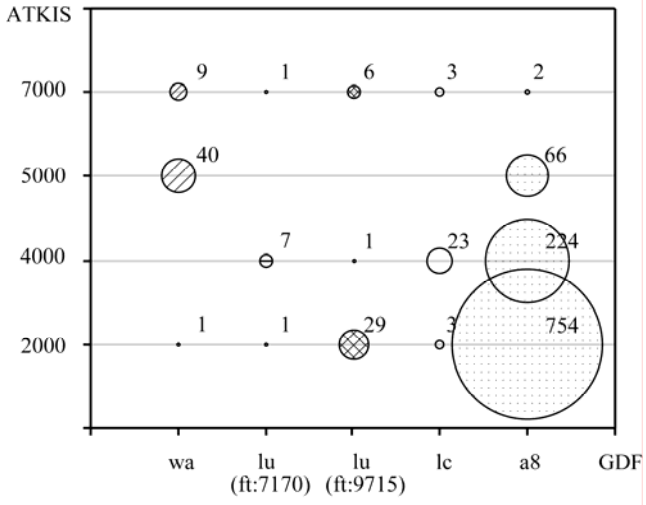

Figure 2: Results from the investigation of 1:n relation with the constraint $R > 80\,\%$ (number of matching candidates are given in the circles).

The results, illustrated in figure 2, show that the intersection between the GDF object group a8 (administrative area) with all ATKIS object groups returns a large number of instances. This indicates that group a8 consists of large objects that contain nearly all the object groups of the other data set. The same holds for the ATKIS objects 7000 (also administrative areas), that also intersects with all GDF object groups. This is an indication that these objects represent overlay or a kind of container objects.

Inspecting the other correspondences reveal dominating relations between single object groups of one data set and single object groups of the other. Thus, semantic similarities in terms of equivalences are visible. These similarities can be seen e.g. in 2000 (urban area, e.g. industrial area) and lu (landuse area) with feature type 9715 (industrial area) or 4000 (vegetation) and lu (landuse area) with featuretyp 7170 (park, garden). Further correspondences are found between 4000 (vegetation) and lc (landcover area, e.g. woodland) and 5000 (water areas) and wa (water areas) as the clearest example.

This is a first step in automatic identification of similar semantic objects through an analysis of geometries. To improve results additional shape describing parameters e.g. compactness or rectangularity have to be introduced to the analysis. Furthermore, also attributes can be included. The goal is an automatic determination of the transformation rules, like the one that is visible in the example above, namely, the semantic equivalence of water objects in both data sets: wa and 5000. This is then the basis for the next step, the geometric adaptation in order to achieve one fused data set.

Geometric data fusion

Based on the given semantic mappings, corresponding objects in both data set can be identified and overlaid. As they typically will not match exactly, a fusion or alignment process is needed in order to harmonize their geometry. Due to the different underlying modelling strategies, there are not only 1:1-relationsships, but also 1:n-relations, which the algorithm has to compensate for. To this end an automated workflow has been implemented which uses the well known ICP algorithm (Besl & McKay, 1992) and a special approach that also allows a partial, distance dependent alignment. The ICP algorithm in our case uses a four parameter transformation for the alignment of two geometries, which results in a coarse fit. The decision, whether the two object instances really match can be taken based on an inspection of the transformation parameters: if the scale parameter is close to 1 and the rotation and translation values are small enough (near zero), one can assume that the objects coarsely fit, otherwise there are larger discrepancies and an geometric adaptation is not possible. In the subsequent step a partial alignment will be performed using an approach based on the identification of point-to-point relations between corresponding objects. Once these 1:1-relations are known – which are only allowed between objects within a given distance – the interpolation of a new common geometry can be performed. This flexibly allows to either fully adapt one data set to the other, or to generate an intermediate geometry by averaging the geometries of the two input objects. In Figure 3 examples for the adjustment are given.

Figure 3: Results from the alignment process: Left - ATKIS (solid lines) and GDF (filled) overlaid, ATKIS and symmetric differences (grey) before alignment (middle), ATKIS and symmetric differences (grey) after alignment.

The result of the adjustment is satisfying: for the objects that are similar, a perfect adaptation could be achieved. In cases, where there are obvious dissimilarities between the objects, also the adaptation was not possible. As a measure for the adaptation quality, the symmetric difference of the two objects after the adjustment is used. In the case of successful 1:1-matches, the measure yields a

value close to 0; in the case of successful 1:n-matches, a non-zero value is returned, as there are gaps between the aggregated objects. Finally, there are also cases, where the alignment was not possible, due to the following reasons: the boundaries of the objects were too far apart (outside the given thresholds), or there were topological differences between the two data sets, so a 1:1-correspondence between the vertices of the two data sets were not found. The symmetric difference between the water-areas before the alignment process has been 134.536,88 m². In total, the measure yielded 60.554,46 m², which is an amelioration of the original correspondences of 45 percent. When neglecting the obvious cases where no perfect alignment was possible, the value reduces to 16.311,71 m². This value still includes the 1:n-relationships that still contain the extra areas between the aggregated objects.

The results of the combined alignment strategy are stored as a collection of displacement-vectors of the individual vertices of the objects. Based on this a vector field can be created on which the alignment of the remaining objects is carried out. Different strategies can be applied accomplish the geometric alignment. Simple alignment can be done by application of a distance weighted method; more sophisticated methods allow the introduction of geometric constraints to the alignment process in order to preserve characteristic object features or relations.

Summary and outlook

The presentation described ongoing work on semantic and geometric data integration. The work on semantic integration is in a very early stage, so in the future, there will be a focus on the research on the automatic instance based determination of transformation rules. Here, association rules and especially hierarchical or multi-level association rules will be investigated (Han & Kamber, 2001). Another important issue is to deal with data sets of different scales or aggregation levels: in this case generalization relations have to be modelled and taken into account, e.g. the fact that polygon objects like rivers can be represented as linear objects in the smaller scale. Another example are road junctions that can be of different complexity depending on the scale (see e.g. Mustière 2006).

BIBLIOGRAPHY

Balley S., Parent C. & Spaccapietra S., 2004: Modelling geographic data with multiple representation. International Journal of Geographical Information Science, 18 (4): 327-352.

Besl P.J. & McKay N.D., 1992: A Method for Registration of 3-D Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Computer Society, (14): 239-256.

Bishr Y., 1997: Semantic Aspects of Interoperable GIS. Wageningen Agricultural University and International Institute for Aerospace Survey and Earth Science (ITC), Enschede.

Doytsher Y., 2000: A rubber sheeting algorithm for non-rectangular maps, Computers Geosciences, Pergamon Press, Inc., 26: 1001-1010.

Duckham, M., Worboys, M.F., 2005: An algebraic approach to automated information fusion. International Journal of Geographical Information Science. v19 n5, 537-557.

Han, J. & M. Kamber: Data Mining: Concepts and Techniques, Morgan Kaufann, 2001.

Hettwer J. & Benning W., 2000: Nachbarschaftstreue Koordinatenberechnung in der Kartenhomogenisierung, Allgemeine Vermessungsnachrichten, 107: 194-197.

Kokla, M., 2006: Guidelines on Geographic Ontology Integration, ISPRS Technical Commission II Symposium, 12-14 July, 2006, Vienna, Austria.

Kuhn W., 2003: Semantic reference systems. International Journal of Geographical Information Science, 17 (5): 405-409.

Mustière, S., 2006. Results of experiments on automated matching of networks at different scales, ISPRS WG II Workshop, Hannover, Germany, 22-24 February, 92-100.

Saalfeld A., 1988: Automated Map Compilation. International Journal of Geographical Information Systems, 2 (3): 217-228.

Sarjakoski, T., Sarjakoski, L. T., Lehto, L., Sester, M., Illert, A., Nissen, F., Rystedt, R. and R. Ruotsalainen , 2002. Geospatial Info-mobility Services - A Challenge for National Mapping Agencies. Proceedings of the Joint International Symposium on "GeoSpatial Theory, Processing and Applications" (ISPRS/Commission IV, SDH2002), Ottawa, Canada, July 8-12, 2002, 5 p, CD-rom.

Stoimenov L. & Djordjevic-Kajan S., 2002: Framework for Semantic GIS Interoperability. FACTA Universitatis, Series Mathematics and Informatics, 17 (2002): 107-125.

Vangenot C., 2004: Multi-representation in spatial database using the MADS conceptual model. ICA Workshop on Generalisation and Multiple representation – 20–21 August 2004, Leicester.

Volz, S., 2005: Data-Driven Matching of Geospatial Schemas. In: Cohn, A.G., Mark, D.M. (eds.): Spatial Information Theory. Proceedings of the International Conference on Spatial Information Theory (COSIT '05), Ellicottville, NY. Lecture Notes in Computer Science 3693, Springer, pp. 115-132 .

Walter V. & Fritsch D., 1999: Matching Spatial Data Sets: a Statistical Approach, Journal of Geographical Information Science, 13(5): 445-473.

Acknowledgement: This research is being funded by the German Science Foundation (DFG) and the German Ministry of Science (BMBF).