Kieler, B.: *A geometry-driven approach for the semantic integration of geodata sets*. In: Proc. of 23rd International Cartographic Conference, 4-10 August 2007, Moscow, Russia.

Birgit Kieler was born in 1979 in Berlin and attended a professional training as land surveying technician in 2001. Subsequently she studied Geodetic Science at the Leibniz University of Hannover and graduated in 2006 obtaining her Master's degree (Dipl.-Ing.). Since June 2006 she is a research assistant of the Institute of Cartography and Geoinformatics at the Leibniz University of Hannover. Her research focuses on semantic integration of data of heterogeneous origin. In a project funded by the German Research Foundation she develops methods for the automatic semantic transformation between geo-ontologies.

A GEOMETRY-DRIVEN APPROACH FOR THE SEMANTIC INTEGRATION OF GEODATA SETS

Birgit Kieler

Institute of Cartography and Geoinformatics, Leibniz University Hannover, Germany

Appelstraße 9A, 30167 Hanover, Germany

birgit.kieler@ikg.uni-hannover.de

ABSTRACT

The paper tackles the problem of data integration and describes an approach for instance-based semantic integration, which improves interoperability and information sharing between data sets of different origin. The approach automatically identifies semantic correspondences between object groups of two different geodata sets by the analysis of geometrical characteristics of the objects. For data sets with the same geographic extent corresponding objects will be detected in a first step by a simple geometric overlay. From these results on instance level the semantic correspondences on class level will be derived. Subsequently, the approach will be extended for the case that geodata sets are not situated within the same area. Therefore a function will be established for a uniform, generic and objective description of subjectively perceived object groups like lakes, woodland or roads. This function contains geometrical and topological characteristics and existing attribute values and allows the identification of similar semantic object groups.

INTRODUCTION

The growing availability of data, especially over the Internet (world wide web) via web services allows for a growing interoperability and combination of data sets of different

origin. This, however, presumes that the content of the data is known in order to draw meaningful and valid conclusions. Thus, for beneficial data integration, both the semantic and the geometric correspondences of these data sets have to be known.

If the semantic relationships between different ontologies are known, a geometric integration for example a fusion or alignment process between the data sets regarding their geometry is possible. However, the normal case is that the semantic relationships are unknown due to the lack of explicit semantics and the poor documentation of these data. Currently, the identification of corresponding semantic object classes is typically done manually, because the process requires expert knowledge based on the precise meaning of the terminology used by the organizations, which capture and model the data sets. Such a manual process is adequate for a limited number of data sets but no longer feasible for an integration of arbitrary data sets that can be loaded from the Internet. Therefore, our goal is to identify corresponding semantic object groups like lakes, woodland or roads in different data sets automatically without expert knowledge, only by the analysis of geometrical and topological characteristics of the object classes itself.

In this paper two cases are discussed. In the first case, we use two geodata sets of similar scale within the same geographical extent. Our approach is to use an instance based determination of semantic transformation rules between their ontologies.

In the second case we try to infer semantic correspondences between data sets that are not within the same area. For the identification of similar semantic objects a function that contains geometrical and topological characteristics of certain object groups is used. Using these instance correspondences a derivation of semantic class correspondences becomes also possible.

The paper is organized as follows. In the next section the background of research in this field is sketched as well as references to existing work are given. Then, our methods for the two integration cases are presented: first the semantic integration of data with similar scale and same extent, then for data of different extent. First results are given introducing some examples. The paper concludes giving some thoughts on future work.

RELATED WORK

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 most difficult problem is the semantic interoperability. The general

approach identifying semantic correspondences is to do schema integration manually using expert knowledge of given object catalogues or ontologies (KOKLA 2006). Such a process is not adequate and not feasible if arbitrary data sets, e.g. downloaded from the Internet have to be integrated. RODRIGUEZ & EGENHOFER (2003) use equality and similarity measures to determine relations between classes from different ontologies. Another method to automate the integration process is a so called instance-based or extensional determination of schema transformation rules (VOLZ 2005, DUCKHAM & WORBOYS 2005). The underlying idea of this approach is that, if two objects have an identical name and / or geometrically coincide, then they probably also have something in common on the semantic level. By contrast FONSECA et al. (2006) presented a framework for measuring the degree of interoperability between geo-ontologies, which only compares the descriptions of the ontologies and not the data themselves. The drawback of this approach is, that in the general case, it can not be assumed, that the names or descriptions of objects in different data sets are the same – except objects with a unique given name like names of cities or roads. Thus using the geometric relations to infer a semantic relation can be more promising. VOLZ (2006) used this approach to link two linear data sets of similar scale.

INTRODUCTION OF TEST DATA

For this work two topographic geodata sets, ATKIS and GDF data, were used: whereas 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 (e.g. 1:25K, 1:50K, 1:250K and 1:1000K), GDF (Geographic Data Files) data was specially developed for purposes of vehicle navigation and is captured for most areas in Western Europe. In our work, we used the most detailed scale of 1:25K that exists for both data sets.

For the determination of semantic transformation rules between the above introduced data sets, the results of a simple geometric overlay of data covering the same geographical extent are used in the first case. For the second case we used data of different extent. Our test areas are located within the urban area of Hanover and have a size of 25 km².

ATKIS as well as GDF model objects using all geometric types: points, lines and polygons. In this analysis only polygon objects from both data sets were used. Tab. 1 lists the object groups and the quantities of objects in the respective groups exemplarily for the first analysis case. 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.

In the analysis process we assume that the data sets are organized in layers. That means, the objects of the data sets are not modelled in a tessellation. From this follows that at one location more than one object class can exist, as illustrated in Fig. 1. This is especially true for administrative boundary objects that often encompass larger areas and are so called "container objects".

data set	object groups		total
*D*1 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
*D*2 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

Tab. 1: Used object groups and quantities from ATKIS and GDF data sets in the test area for the first case

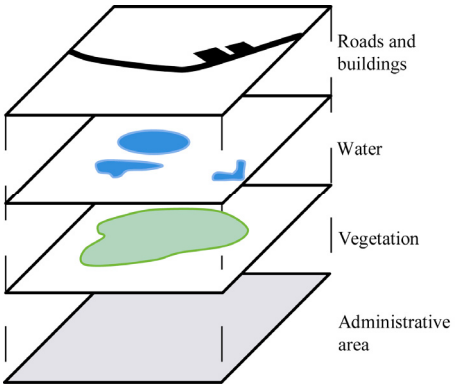

Fig. 1: Simplified representation of the layer structure

METHOD OF THE GEOMETRIC OVERLAY

For the identification of corresponding objects a simple geometric overlay of the data sets representing the same geometric extent is done. Due to the data organization in layers an intersection of the data sets returns more than one matching candidate which could cause difficulties for the further analysis. But not only the layer structure, also the geometric

discrepancies at the object boundaries themselves, causes an increasing number of possible matching candidates, because the adjacent objects will be taken into account. To restrict this number of correspondences and to improve the results the additional geometric criterion **area**, especially the intersection area *I* and object size *O*, is introduced in the analysis.

To accomplish finding corresponding objects all objects of data set *D*1 (ATKIS) have to be intersected with all layers of data set *D*2 (GDF) and vice versa. In this process the overlay ratios of the intersection area and the size of the objects as illustrated in Fig. 2 and Eq. 1 are analysed to get further information about possible matching partners.

$$R_{O,1} = \frac{I \cdot 100\%}{O_1} \text{ and } R_{O,2} = \frac{I \cdot 100\%}{O_2} \qquad \text{Eq. (1)}$$

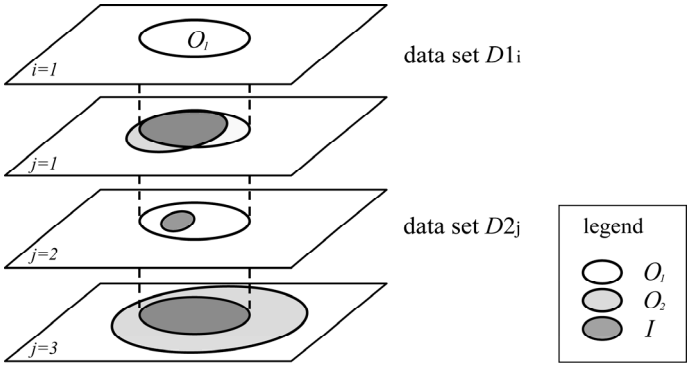

Fig. 2: Different ratios between the intersection area I and object sizes O. *j=1* $R_{O,1} \geq 80\% \wedge R_{O,2} \geq 80\%$; *j=2* $R_{O,2} = 100\%$; *j=3* $R_{O,1} = 100\%$

Taking small geometric differences into account, we consider objects as matching, when the ratio is 80% or better. In the case of a 1:1 relationship the following condition has to hold: $R_{O,1} \geq 80\% \wedge R_{O,2} \geq 80\%$. But if the data sets are differently modelled and as described above contain more aggregated objects than the other data set, the search for 1:1 relations returns only few matching candidates. In our example only seven objects meet the condition. In Fig. 3 some examples of such relations of selected object classes are presented.

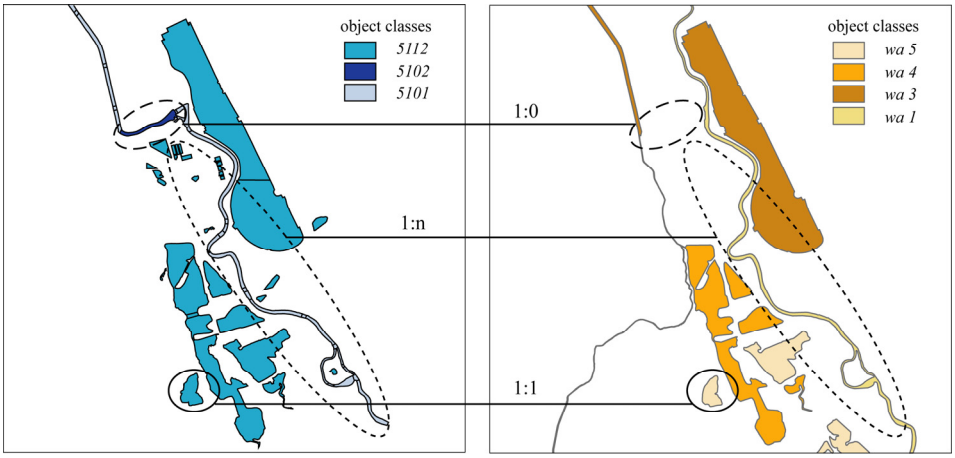

Fig. 3: Data set D1-ATKIS (left) and D2-GDF (right) and object relations 1:0, 1:1, 1:n for selected object classes

Therefore the analysis is extended from 1:1 relations to 1:n relations. In that case the condition $R_{O,1} \geq 80\% \vee R_{O,2} \geq 80\%$ must be kept. The results of this overlay and the area comparison method on instance level can be presented for each relation in a frequency matrix that displays the number of possible matching candidates, as illustrated in Fig. 4.

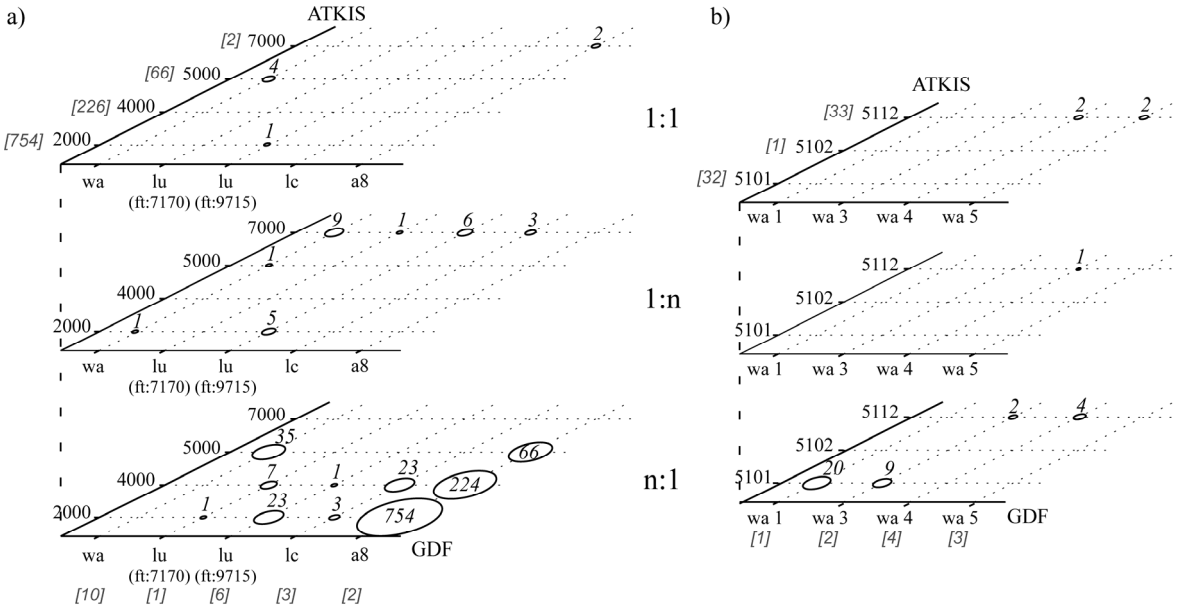

Fig.4: Frequency matrices for 1:1 (top), 1:n (middle) and n:1 (bottom) relations of data sets *D*1 (ATKIS) and *D*2 (GDF). The meaning of the abbreviations can be taken from Tab.1. The total numbers of the object domains are displayed in grey brackets. In Fig. 4a) the entire results and in b) the detailed results of object domains *5000* (ATKIS) and *wa* (GDF) are shown.

Using these results of the instance level the correspondences on class level can subsequently be inferred. Besides the very obvious 1:0-disjunction relations ($D1_i \cap D2_j = \varnothing$) that are detected between object domain *wa* (water element area) and *4000* (vegetation) respectively *lc* (landcover area), *lu* (landuse area) and *5000* (water area), also a 1:1-equivalence relation ($D1_i \equiv D2_j$) between the object domains *7000* (region) and *a8* (administrative area) is detected, because all objects of both object domains meet the condition $R_{O,1} \geq 80\% \wedge R_{O,2} \geq 80\%$. Inspecting the other correspondences a great amount of 1:n-inclusion relations ($D1_i \subseteq D2_j$) is additionally found (see in Fig. 4a) middle and bottom). But not all of these relations represent true 1:n relations, as represented in the examples in Fig. 3, because all relations to container objects, which are large objects and contain nearly all other objects of the other data sets, are included. These relations to container objects have to be dismissed from the further analysis. In our example the objects of the ATKIS object domain *7000* and GDF object domain *a8* are identified as such container objects.

Between object domain *5000* and *wa* all kinds of relations exist, which is an indicator for semantic relations between the object classes. For a better analysis of the relations in Fig. 4b) the detailed results of the relations are shown. From this results can be inferred that the classes *wa 1* and *wa 3* match to object class *5101* (stream, river, ditch) and the classes *wa 3*, *wa 4* and *wa 5* match to object class *5112* (lake, dam, pond) as illustrated in Fig. 5. Object class *5102* (channel) has no matching partners at all in this examination.

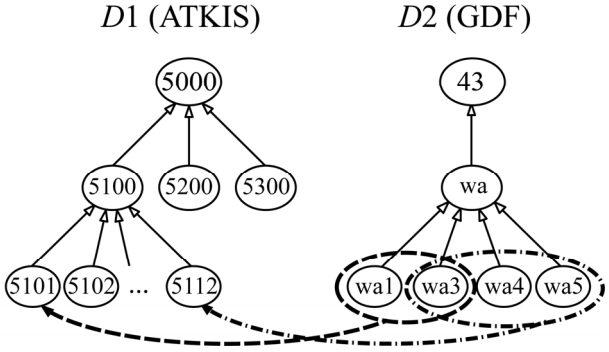

Fig.5: Semantic correspondences on detailed class level between *D*1 (ATKIS) and *D*2 (GDF)

The object class *wa 3* is being matched to *5101* and also *5112*. Inspecting the instances of *wa 3* more in detail (see Fig. 3) reveals that the class contains two very shape-differing objects, i.e. a river object and a lake object. An expert would assign the lake object to *5101*

and river object to *5112* with high probability. In order to automate this expert decision, additional geometrical characteristics must be included into the analysis process extending the intersection area criterion regarded so far. Which geometrical characteristics are possible to use and how they possibly improve the process will be described in the next section.

MODELLING INTRINSIC CHARACTERISTICS OF FEATURES

Because the data sets are not always existent for the same spatial extent, the geometric overlay procedure to identify similar semantic object groups is not possible and therefore the matching task is more complicated. A first idea to solve this problem is to establish generic object models, which describe the intrinsic characteristics of particular object groups like lakes, rivers, buildings or roads in a generic way. Using this model the predefined semantic object groups can be detected automatically within the new data sets. In addition, with the knowledge about the characteristics of specific object types, the results of the geometric overlay in the above described first case could also be improved. Apart from characteristics of object geometry G also existing attribute values AV and the topology T to neighbour objects will be used to create a abstract feature model. But which shape descriptive parameters are available in order to describe the geometry of an object precisely and nevertheless objectively, and thus possess the potential, to make a clear identification of particular types of objects possible? Besides the parameter object area O also perimeter P and length of an object are of great importance. Basing upon these three basic characteristics further parameters can be derived, that allow to draw conclusions on the actual shape of an object. Additionally to the elongation E, that presents the relationship of the two principal axes to each other, the compactness C, as the relationship of perimeter and area, and the rectangularity R, a measure of the angle deviations along the object outline, give inferences about important geometrical characteristics. Also different kinds of hulls, e.g. the minimum bounding rectangle MBR or the convex hull CH supply important decision support for the assignment. To model the perception of objects as a function of these characteristics is a big challenge. In Fig. 6 examples are displayed, which can be characterized, using the following properties. Lakes and ponds do have in most cases a round, organic outline, whereas rivers show a lengthier but nonetheless organic shape with partial parallel borders. On the other hand a road network consists of a netlike structure with mostly rectangular junctions and buildings are of a small size and have generally parallel borders and right angles. Generic object models have to be established

combining these shape-describing parameters to express the different perception of these objects types.

Fig.6: Selected types of objects which have to be described with the object-describing functions. From left to right: lakes respectively ponds, rivers, a street network and buildings.

In order to produce most exact functions it is necessary to compute these parameters of real world objects, such as the objects of the first case in a first step. In Tab.2 the results of the parameter analysis are arranged in terms of mean values. For a better comparison also the results of aggregated object classes labelled with *agg*, where directly neighbouring objects are merged, are listed.

object class	total	***O*** *[m²]*	***P*** *[m]*	***E***	***C***	***R***	***MBR***		R_c = *O/MBR*
							length [m]	*width [m]*	
wa 1	1	198.057,68	14.944,02	6,65	0,01	0,14	5.446,02	818,63	0,04
wa 3	2	432.464,23	9.310,58	4,02	0,15	0,25	3.701,04	944,02	0,30
wa 4	4	128.669,67	2.446,04	2,74	0,39	0,24	794,78	248,35	0,52
wa 5	3	111.743,97	2.222,65	1,47	0,39	0,25	523,54	362,50	0,57
5101	32	7.035,17	511,95	3,24	0,50	0,38	205,07	61,20	0,68
5101 agg	2	112.562,71	7216,06	8,54	0,04	0,10	2.717,96	416,44	0,14
5102	1	21.500,61	1.314,17	5,85	0,16	0,22	564,51	96,49	0,40
5112	33	48.170,12	809,57	2,15	0,55	0,39	273,32	129,63	0,70
5112 agg	26	61.138099	982,84	0,95	0,52	0,40	336,29	129,20	0,67

Tab. 2: Mean values for shape describing parameters of selected object groups of the first case test area

A closer look at the results indicates some specific object characteristics, e.g. rivers (*wa 1*, *5101 agg [light grey]*) have a small compactness and rectangularity as well as a small ratio R_c between the object area *O* and the area of the minimum bounding rectangle *MBR,* but compared with the other values a large elongation. In contrast lakes and ponds (*wa 4*, *wa 5, 5112 agg [dark grey]*) show a small elongation and relatively large values for compactness

and R_c. Exemplarily for the object types *river* and *lake* the following models with ranges for some parameters were established.

$$river = \{ obj \mid C(obj) \leq 0{,}10 \wedge R(obj) \leq 0{,}15 \wedge R_c(obj) \leq 0{,}25 \wedge E(obj) \geq 5{,}5 \}$$

$$lake = \{ obj \mid C(obj) \geq 0{,}35 \wedge R(obj) \geq 0{,}20 \wedge R_c(obj) \geq 0{,}50 \wedge E(obj) \leq 3{,}0 \}$$

It should be pointed out that these functions have no general validity, because they were derived only from a small number of examples. In order to increase the probability for a correct retrieval and identification of objects, also more parameters must be evaluated statistically from a sufficiently large data base apart from improved ranges of values of the examined parameters. Additionally attribute values like proper names and topological relations to neighbouring objects, i.e. that lakes are generally surrounded by vegetation areas, are introduced in the further analysis. The quality of the semantic integration results will be improved combining different parameters and more detailed modelling of the different intrinsic characteristics of object types.

CONCLUSION AND FUTURE WORK

The paper describes ongoing work on semantic data interoperability. Automatic procedure for integrating data sets of different origin are very important for reusing geodata. Currently our work on semantic integration is in a very early stage. In the future, we will be a focus mainly on improving the identification of similar semantic objects by an extended analysis, especially by automating this process. 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 are represented as linear objects in a smaller scale.

ACKNOWLEDGEMENTS

This work is part of the project “Automatic Semantic Transformation between Geo-Ontologies” and is funded by the German Research Foundation (DFG). It is part of the bilateral bundle-project entitled “Interoperation of 3D Urban Geoinformation” in cooperation with China.

REFERENCES

BISHR, Y., 1997: Semantic aspects of interoperable GIS. – Wageningen Agricultural University and International Institute for Aerospace Survey and and Earth Science (ITC), Enschede.

DUCKHAM, M. & WORBOYS, M., 2005: An algebraic approach to automated geospatial information fusion. – International Journal of Geographical Science **19** (5): 537–557.

FONSECA, F., CÂMARA, G. & MONTEIRO, A.M., 2006: A Framework for Measuring the Interoperability of Geo-Ontologies. – Spatial Cognition and Computation **6** (4): 309–331, Lawrence Erlbaum Associates, Inc.

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

RODRIGUEZ, A. & EGENHOFER, M., 2003: Determining Semantic Similarities Among entity classes from different ontologies. – IEEE Transactions on Knowledge and Data Engineering, **12** (2): 442–456.

VOLZ, S., 2005: Data-Driven Matching of Geospatial Schemas. – Cohn, A.G., Mark, D.M. (eds.): Spatial Information Theory. Proceedings of the International Conference on Spatial

VOLZ, S., 2006: Modellierung und Nutzung von Relationen zwischen Mehrfachrepräsentationen in GeoInformationssystemen. – Dissertation, Fakultät Luft- und Raumfahrttechnik und Geodäsie der Universität Stuttgart, Stuttgart.