LINKING DIFFERENT GEOSPATIAL DATABASES BY EXPLICIT RELATIONS

Steffen Volz *, Volker Walter

University of Stuttgart, Institute for Photogrammetry, Geschwister-Scholl-Str. 24D, 70174 Stuttgart, Germany - (steffen.volz, volker.walter)@ifp.uni-stuttgart.de

Commission IV WG IV/2

KEY WORDS: GIS, Integration, Interoperability, Open Systems, Federated, Spatial Infrastructures

ABSTRACT:

The growing awareness of the importance of spatial information has led to a continuously increasing demand for geospatial data. Thus, a lot of different companies and institutions have evolved that are aiming at satisfying this demand by capturing the real world according to the needs of different applications. This process involves that one and the same real world object is stored in several representations in different geospatial databases. The integration of these representations is a major research challenge in the field of GIS. In our approach we use a semi-automatic matching tool for defining relations between representations on the instance level. These relations are used in order to automatically derive an integrated data schema. The relations on the instance level as well as the integrated data schema can be used for automatically merging data sets, for automatically transferring updates from one data set into another or for a common analysis of spatial data from different sources.

1. INTRODUCTION

Many organisations have made major investments in capturing spatial data. However, an exchange of these datasets or a combined use is done only very rarely. This situation is going to be changed by using new internet technologies. One of the driving forces in this field has been the development of Web Services for Geographical Information Systems. Web Services are interoperable applications that are accessible with standardized interfaces in the internet. With that kind of technology it is possible to write platform independent programs which can be shared between different users. Therefore, Web Services could be seen as the basic technology for interoperable Geographical Information Systems. Web Services are also the basis for the realisation of Spatial Data Infrastructures. The aims of these infrastructures consist of improving the quality of data, reducing costs, making data more accessible and providing consistent datasets.

However, very often, one main important aspect is ignored here. It is not sufficient to provide only platform independent programs to realize Spatial Data Infrastructures; it is also necessary to provide integration techniques for spatial data. Because the data are captured by different organisations, one object of the landscape is stored in several databases in different data models, at different acquisition times, with different quality characteristics or in different scales.

The integration of spatial data can be done at two levels: (1) an integration of the spatial instances or (2) an integration of the data schemas. Research in the spatial domain was mainly focusing either on matching techniques on the instance level or on ontology based approaches on the schema level. Up to now, no approaches could be found which are combining the two paradigms. In this paper we examine how the integration of different data schemas can be derived automatically from integrated data instances (see figure 1).

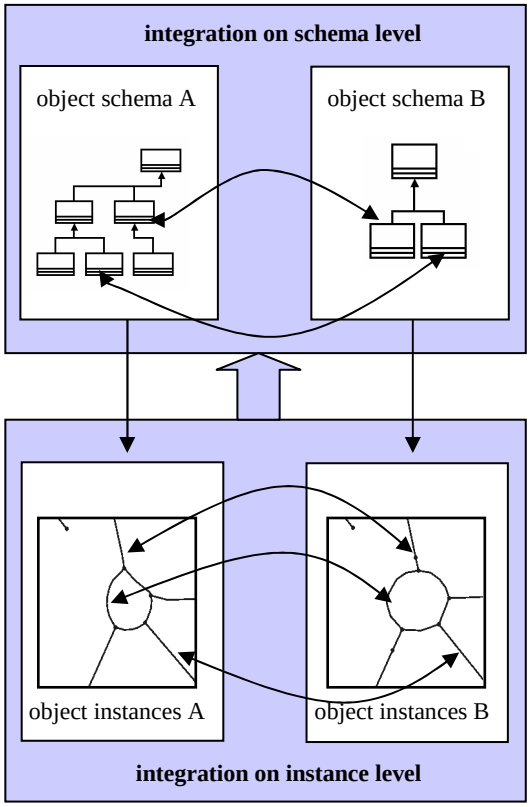

Figure 1. Basic idea: deriving schema relations by analyzing instance relations.

The integration of data instances is typically done with matching techniques. There already exists a lot of research on how this matching can be done in an automatic way, but at the moment there is no common approach, because this problem is very application dependent. Therefore we use a semi-automatic

* Corresponding author

matching tool for our research that gives us the possibility to match the data comfortably by hand and use these matchings as an input for the integration of data schemas.

This work is part of the Nexus project (Nexus 04). In the Nexus project, we are developing an open platform for all possible types of mobile, location-based information systems. In order to realize a generic approach in Nexus, different data providers have to be able to integrate their data into the Nexus world model. For this reason, a schema integration takes place that maps the object classes of existing data schemas from data providers onto the classes of the Nexus schema. At the moment this process is done manually. In this paper we show how this can be done in an automatic way. The paper first gives an overview on related work. In section 3, it is discussed how spatial databases can be related. Section 4 comprises a detailed explanation of the realization of our approach.

2. RELATED WORK

The topic of spatial data integration is very much related to the research areas listed below. Some of their aspects will be briefly presented in the following sections.

The notion of the research presented in this paper has already been addressed by (Uitermark 1996): "Geographic Data set integration (or map integration) is the process of establishing relationships between corresponding object instances in different, autonomously produced, geographic data sets of a certain region. The purpose of geographic data set integration is to share information between different geographic information sources".

2.1 Matching and conflation

Concerning the matching of spatial objects, the basic idea is to express and to evaluate the similarity of spatial features. If a certain degree of similarity can be detected, two features can be assigned to each other. (Bruns & Egenhofer 1996) have adopted this basic assumption and count the steps that have to be taken to transform one representation into another representation. The number of steps can then be interpreted as a similarity measure.

A fundamental, line-based matching approach for street network data has been presented by (Walter and Fritsch 1999). In a first step, the algorithm finds all potential correspondencies of topologically connected line elements in two source data sets by performing a buffer operation. The matching candidates are stored in a list. This list is ambiguous and typically contains a large amount of *n:m* matching pairs. Then, unlikely matching pairs are identified and eliminated using relational parameters like topologic information and feature-based parameters like line angles. The result is a smaller but still ambiguous list with potential matching pairs. These matching pairs are evaluated with a merit function in order to compute a unique combination of matching pairs which represents the solution of the matching problem. This is a combinatorial problem which is solved with an A* algorithm.

A point-based matching method was proposed, for example, in (Bofinger 2001). The algorithm developed here is based on the idea of describing intersections of streets, i.e. nodes of a street network, by an explicitly defined code. The code consists of point coordinates, abbreviations and names of incident streets and the number of linked edges. For each intersection, such a code is created. By comparing the codes of the intersections within different data sets and by assigning the intersections with the most similar codes to each other, references can be derived.

The problem of conflation is for example being tackled by (Cobb et al. 1998). The merging process is defined here as "feature deconfliction", where all parts of a matched feature pair are unified into a single "better'" feature. The conflation algorithm has to decide, which properties are preserved in the resulting instance. In their approach, the authors are also taking into account the data quality information of the corresponding instances.

2.2 Semantic data integration and ontologies

According to (Uitermark et al. 1999), semantic integration can be understood as a communication process since two partners who want to communicate have to have the same understanding of the objects they are talking about.

In the database domain, some work has been done regarding schema matching by (Do and Rahm 2002), where schemas are compared using parameters like element names, data types or further structural information. In the field of GIS, a lot of different approaches have been carried out using ontologies. Ontologies can be defined as formalized specifications of concepts about objects of the real world from a certain application perspective (Gruber 93). Whereas database schemas require a digital representation, ontologies are just abstract views on the semantics of things. There is only one ontology for an object in a certain application domain, but there can be multiple database representations (Fonseca et al. 2002). Consequently, concerning schema integration, two cases have to be considered (Hakimpour and Timpf 2001):

1. Database schemas are based on the same ontology: only synonyms and homonyms have to be detected to perform an integration.
2. Database schemas are based on different ontologies (from different application domains): a common ontology has to be created by detecting the similarities of the source ontologies.

The authors are presenting a formalism for the representation of ontologies, the so-called Description Logic (DL). Each user community can define its perception of an object using DL and then different ontologies can be merged.

Another example on how to integrate different semantics of spatial data is provided by (Bishr et al. 1999). The approach consists of two components, the Semantic Wrapper and the Semantic Mapper. Objects of different spatial databases are wrapped by the Semantic Wrapper and have to conform to a predefined interface so that they can be recognised by the Semantic Mapper. This interface is specific for a certain application domain like transportation, topography, etc. On the level of the Semantic Mapper, the semantics of two objects can be compared and the schematic and semantic differences between them can be resolved.

2.3 Standardization

The question of interoperability of GIS is mainly addressed by the OpenGIS Consortium (OGC 2004) and the Technical Commission 211 of the International Standards Organization (ISO-TC211 2004). Both institutions are closely linked.

Traditionally, their work is focusing on the technical part of interoperability, i.e. on the specification of data and component interfaces. With the abstract specification concerning semantics and information communities (OGC 1999), the OGC published first ideas concerning the realization of semantic interoperability. In this document, a system has been proposed to reduce the loss of information while transferring data from one user community to the other one. However, up to now no steps have been taken to implement the concept.

3. LINKING AND RELATING MULTIPLE REPRESENTATION DATABASES

In this section it is outlined on which levels spatial databases containing multiple representations can be linked. Then, the links or relations, respectively, that can exist on the instance level are investigated and some tasks and problems that have to be dealt with in this context are identified.

3.1 Levels of integration

The integration of different spatial databases can be performed on different levels, as it can be depicted from figure 2.

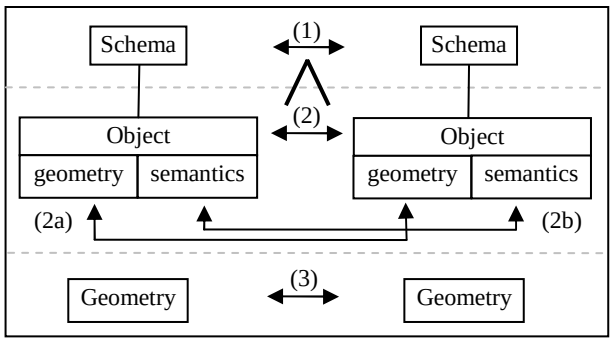

Figure 2. Integration of spatial databases on different levels.

On the one hand, the different object classes (and attributes) of the source schemas can be linked (1). On the other hand, the object instances themselves can be matched by looking at their geometric (including topologic) (2a) and/or semantic properties (2b). Finally, links could only be set up between the geometries (3) of two data sets. Figure 3 shows that there can be differences between the results of (2) and (3).

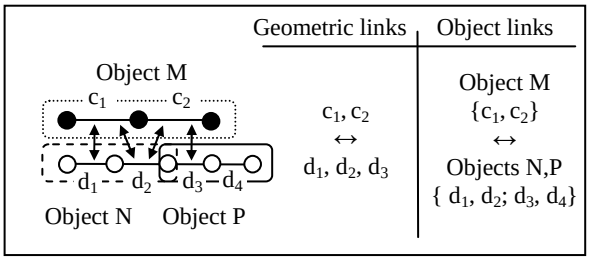

Figure 3. Geometric linking and object linking can lead to different results.

3.2 Links between instances

In our approach, we want to use the results of the integration on the object level in order to derive an integration on the schema level. Therefore, we first have to figure out how instances can be linked or related. Some of the tasks and problems that have to be faced here are outlined in the following sections.

3.2.1 The MultirepresentationalRelation: Basically, links could only be represented as a list containing matching pairs (see figure 3). However, the links between coinciding representations can be further described. In our case, this is necessary because we want to have information about the degree of similarity of corresponding instances. For this reason, we introduced a so-called MultirepresentationalRelation object to connect multiple representations. Within such an object, all information on how representations are related can be stored. Its attributes contain the general, geometric, topologic and semantic/meta properties of the representations taking part. Some of them are listed below:

General attributes:

- The IDs of the corresponding objects.
- The cardinality of the relation.

Geometric attributes:

- Geometric types of corresponding features (see figure 4).

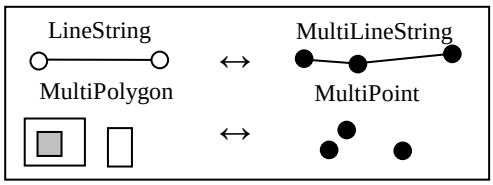

Figure 4. Features of different geometric types can take part in a relation.

- Geometric resolution and position accuracy of corresponding objects (e.g. one geometry has been captured in a large scale of 1:1000, another one in a smaller scale of 1:25 000).
- Geometric comparisons of corresponding features, depending on the geometric shape (e.g. angles, distance measures between objects, comparisons of their area, their length, etc.; see figure 6).

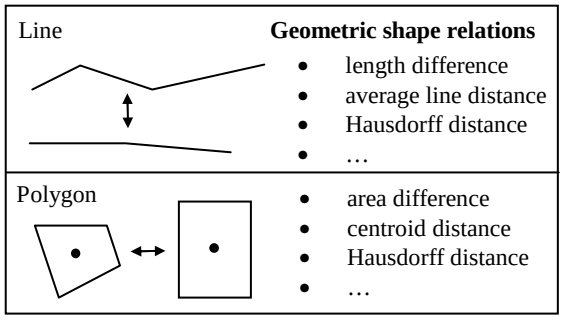

Figure 6. Various measures to compare the similarity of geometric shapes are available.

Topologic attributes:

- Number of adjacent or incident features of corresponding objects.
- Their graph-based topology indicators like e.g. the reachability or eccentricity of nodes that constitute an edge, etc.

Semantic/meta attributes:

- Affiliation of corresponding features to an object class in their source data sets
- Number of corresponding attributes or attribute values of corresponding features.
- metadata like information about the spatial reference systems or data quality parameters, e.g. the means of

acquisition, date of acquisition, etc., of corresponding features

3.2.2 Deriving similarity measures: The result of an integration of corresponding objects is the more significant and useful, the clearer and the more reliable the similarity of the features can be assessed. If good similarity measures are available, then also the applications which are using the results of an integration process, namely the conflation, analysis and update of corresponding instances, can be optimized. In our application, we need similarity measures in order to introduce thresholds. These thresholds shall be used to figure out which degree of similarity we actually need between instances if we want to deduce information about correspondencies between schemas.

A lot of attributes within a MultirepresentationalRelation object can also be interpreted as indicators showing the similarity of related representations, e.g. the geometric distance, the number of adjacent features or the number of corresponding attributes, etc. The task is now to figure out how one global similarity measure (GSM) can be calculated from all the individual similarity measures (ISM). In a first basic approach, we're using a weighted sum:

$$GSM = \sum_{i=0}^{i} ISM_i * weight_i$$

As it has been proposed in (Walter and Fritsch 1999), a statistical approach in order to exploit combinations of measures could be applied as well.

3.2.3 Difficulties in instance matching: When a matching of corresponding instances is performed, we can have simple, non-ambiguous cases of cardinality *1:1, 1:n* or *n:m*. But the process can also involve severe difficulties: cases can occur in which features of different object classes or with different attributes or attribute values are taking part in a *1:n* or an *n:m* relation. Thus, we have "pure" relations, but we can also have "impure" relations (see figure 7).

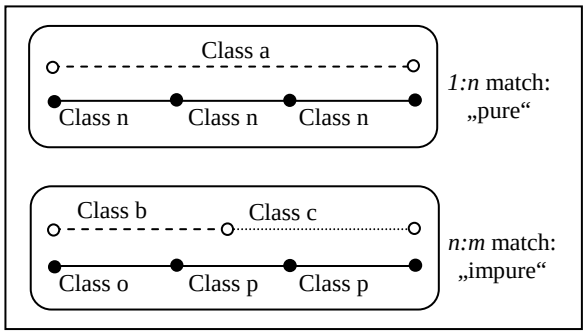

Figure 7. Impure and pure relations between instances.

Impure relations between corresponding representations can constrain the usefulness of our approach since they provoke ambiguities. For this reason, they have to be dealt with appropriately when we infer the correlation between object classes or attributes. Pure relations have to have more influence than impure. Furthermore, measures to assess the degree of impurity have to be found. This is part of our future work.

4. BUILDING AND ANALYZING RELATIONS BETWEEN MULTIPLE REPRESENTATIONS

In the first phase of this research, a tool has been developed that allows building up relations between multiple representations in a semiautomatic way. Once the relations are created they can be used to automatically derive similarity measures for the schemas of the source data sets. This second step is still work in progress, only some first results are available.

The whole software that has been implemented is integrated into an open, Java-based software environment, which has been developed by the Jump project (JUMP 2004). It consists of three modules (in the Jump terminology, they are called plug-ins): the Relation Builder module allows to build up relations between corresponding instances (see figure 9), the Relation Viewer module allows to display these relations and the Relation Analyzer module allows to interpret the relations.

4.1 Building relations

The first step of our approach consists of generating the relations between multiple representations stemming from heterogeneous sources. Basically, it would be optimal to realize this automatically. However, we are not focusing on the automatic creation of relations, but we want to exploit the relations in order to deduce information about schema correspondencies. Therefore, we have realized a semiautomatic approach, where an operator selects corresponding instances in the map view. Involving a human operator can cause inconsistencies, since two operators might interpret a spatial scene differently. Thus, a catalogue of instructions on how to deal with certain situations had to be set up in order to achieve at least similar and comparable results. For example, a rule has to be provided for the matching of street network data that says if topologically separated objects can take part in a *1:n* or an *n:m* relation. In our case, this is possible (see figure 8).

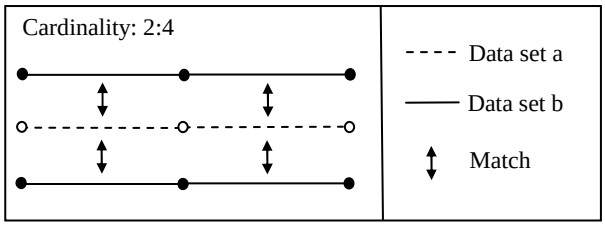

Figure 8. In our case, n:m relations can also be set up between separated road segments.

Up to now, relations have been set up for a test area in the inner city of Stuttgart, covering an area of approximately one square kilometre. It contains street data of Geographic Data Files (GDF) and the Authoritative Topographic Cartographic Information System (ATKIS). GDF is mainly used for car navigation purposes, whereas ATKIS is a topographic database that was set up with the intention to provide spatial data for different kinds of applications. Figure 9 shows a clipping of the test scenario.

Figure 9. Clipping of the test scenario from the inner city of Stuttgart (ATKIS in dotted, GDF in straight lines).

In order to build MultirepresentationalRelation objects for corresponding representations of the two source data sets, the Relation Builder was developed (see figure 10).

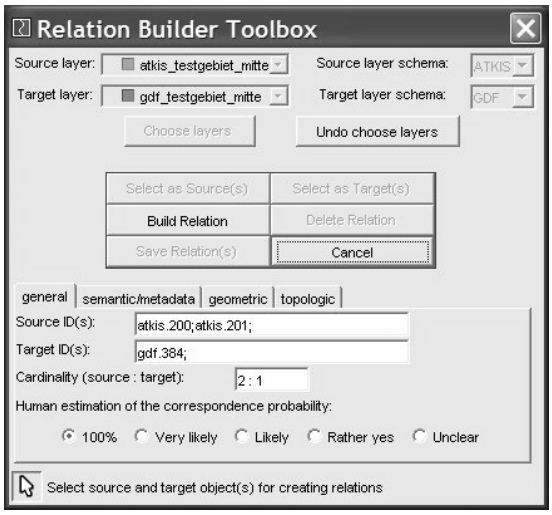

Figure 10. The Relation Builder module.

It provides a feature to select the representations in the map view. The operator can also specify his personal impression about the likelihood of correspondence. Pressing the "Build Relation" button automatically creates the relation with all its attributes for the selected features and greys them out so the operator can see for which objects relations have been set up already.

If all necessary relations have been created, they can be saved in an XML-based MRRL (**M**ulti**R**epresentational **R**elation **L**anguage) file. This file stores general information about the source data sets and all the relations that exist between their individual instances. Figure 10 shows a short extract of this file. The implementation has been done using the (JDOM 2004) library for XML processing in Java.

```
<mreprelation>
    <mreprelation_id>4</mreprelation_id>
        <general_atts>
            <source_ids>atkis.178;atkis.179;</source_ids>
            <target_ids>gdf.363;</target_ids>
            <cardinality>2:1</cardinality>
            <human_estimation>100%</human_estimation>
        </general_atts >
        <semantic_atts>
            <source_classes>Street;Street;</source_classes>
            <target_classes>RDEL;</target_classes>
            <…>…<…/>
        </semantic_atts>
        <geometric_atts>
            <length_difference>-8.78597077</length_difference>
            <…>…<…/>
        </geometric_atts>
        <topologic_atts/>
</mreprelation>
```

Figure 10. Structure of a MultirepresentationalRelation in the MRRL format.

4.2 Analyzing relations

In the work that has been done up to now, we have investigated the correspondencies on the object class level between ATKIS and GDF by analyzing the instance relations of our test area. In first results, we could show, for example, that there are significant correspondencies between the object classes "Road" (RD) and "Intersection" (ISEC) of GDF and "Lane" (LN) of ATKIS, as it is illustrated in figure 11.

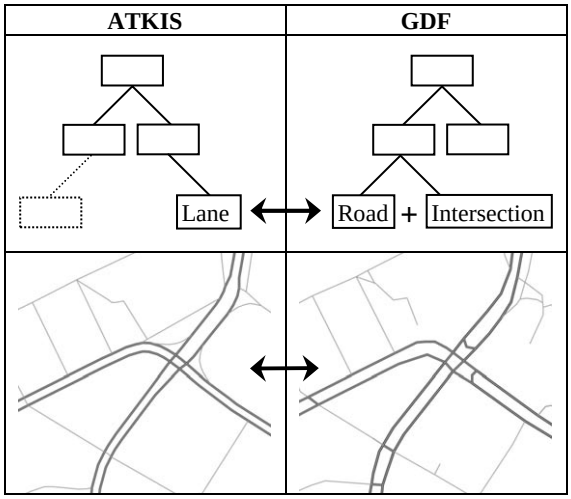

Figure 11. ATKIS Lane (LN) and GDF Road (RD) and Intersection (ISEC) objects show a high degree of similarity.

Within our test data sets, there are 370 features of GDF and 267 of ATKIS. 175 of the 370 GDF instances are either belonging to the RD or the ISEC class. On the other hand, 77 of the 267 ATKIS features are "Lane" objects. We created 57 MultirepresentationalRelation objects for the representations of the investigated classes. Only 4 GDF representations which held RD or ISEC objects couldn't be assigned to LN objects of ATKIS at all (see * in table 1). As an example, we found 4 RD features which were corresponding with a "Tunnel" and a "Street" feature of ATKIS. There were 46 pure relations between the object classes, though, and 4 of the impure matches also supported our hypothesis of a direct correspondence of ATKIS LN and GDF RD and ISEC. In 3 cases, we found only a

partial correspondence, e.g. one ATKIS LN was assigned to one GDF RD and one GDF Road Element (RDEL). So altogether, 7% of the relations are in contradiction to the statement that there are correspondencies between the object classes under investigation, 5.3% do not clearly support our conclusion but 87.7% do speak well for a clear link. Details about the relations between the representations can be depicted from table 1:

	GDF	ATKIS
- 1:n relations	*4*	*19*
- 1:1 relations	*20*	
- n:m relations	*14*	
- impure relations	**11**	
- LN and (ISEC and RD)	4	
- Other classes involved	3	
- Only other classes	4*	
- pure relations	**46**	
- LN and ISEC	13	
- LN and RD	33	

Table 1. Results from the matching of the test area.

It has to be pointed out that these are first results on a small test area. In the near future, the approach has to be verified using larger data sets. Moreover, we are working on combining attributes and object classes in order to detect correspondencies. For example, we expect to have similarities between "Way" objects and "Street" objects with attribute "road_type" = "Community Street" from ATKIS and GDF Road Elements with attribute "functional class = 5".

5. CONCLUSIONS AND OUTLOOK

In this paper we have shown that spatial databases can be linked on different levels. It was our goal to prove that explicit relations on the instance level can be used to derive links on the schema level. First results have been achieved which have to be verified in the future. Furthermore, it is planned to exploit the relations we have set up between instances in order to optimize the processes of conflation, update and analysis of multiple representations. But this is not trivial, especially in the case of *n:m* matches.

6. REFERENCES

Bishr, Y. A., Pundt, H. Rüther, C., 1999. Proceeding on the Road of Semantic Interoperability - Design of a Semantic Mapper based on a Case Study from Transportation, *in*: Proceedings of the 2nd International Conference on Interoperating Geographic Information Systems, Zurich, Lecture Notes in Computer Science, Heidelberg, Berlin, pp. 203-215.

Bofinger, J.M., 2001. Analyse und Implementierung eines Verfahrens zur Referenzierung geographischer Objekte. Diploma Thesis at the Institute for Photogrammetry, University of Stuttgart, unpublished, 76 pages.

Bruns, H. and Egenhofer, M., 1996. Similarity of Spatial Scenes, Seventh International Symposium on Spatial Data Handling (SDH '96), Delft, The Netherlands, pp. 173-184.

Cobb, M., Chung, M., Miller, V., Foley, H., Petry, F., Shaw, K., 1998. A rule-based approach for the conflation of attributed vector data, *GeoInformatica* 2(1), pp. 7-35.

Do, H.H. and Rahm, E., 2002. COMA – A System for Flexible Combination of Schema Matching Approaches, in: Proceedings of the 28th Intl. Conference on Very Large Databases (VLDB) Hongkong, http://www.vldb.org/conf/2002/S17P03.pdf (acc. 26th November 2003), 12 pages.

Fonseca, F., Egenhofer, M., Agouris, P. and Câmara, G., 2002. Using Ontologies for Integrated Geographic Information Systems, *Transactions in GIS* 6(3), pp. 231-257.

Gruber, T., 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 2(5), pp. 199-220.

Hakimpour, F. and Timpf, S., 2001. Using Ontologies for Resolution of Semantic Heterogeneity in GIS; Proc. Of the 4th AGILE Conference on Geographic Information Science, Brno. http://www.ifi.unizh.ch/dbtg/Projects/MIGI/publication/agile2001.pdf (acc. 15th August 2003), 12 pages.

ISO-TC211, 2004: http://www.isotc211.org/ (acc. 26th April 2004).

JDOM, 2004. http://www.jdom.org/ (acc. 15th March 2004)

JUMP, 2004. http://www.jump-project.org/ (acc. 15th March 2004).

Nexus, 2004. http://www.nexus.uni-stuttgart.de/ (acc. 30th April 2004).

OGC 1999. The OpenGIS™ Abstract Specification, Topic 14: Semantics and Information Communities, Version 4. http://www.opengis.org/docs/99-114.pdf (acc. 5th April 2004).

OGC 2004. http://www.opengis.org/ (acc. 20th March 2004).

Uitermark, H., 1996. The integration of geographic databases. Realising geodata interoperability through the hypermap metaphor and a mediator architecture, *in*: Rumor, M., McMillan, R. and Ottens, H. F., Proceedings of the 2nd Joint European Conference & Exhibition on Geographical Information (JEC-GI) '96, Vol. I, Barcelona, pp. 92-95.

Uitermark, H., Vogels, A., van Oosterom, P., 1999. Semantic and geometric aspects of integrating road networks, *in*: Proceedings of the 2nd International Conference on Interoperating Geographic Information Systems, Zurich, Lecture Notes in Computer Science, Springer-Verlag, Heidelberg, Berlin, pp. 177-188.

Walter, V. and Fritsch, D., 1999. Matching Spatial Data Sets: a Statistical Approach, *International Journal of Geographical Information Science* 13(5), pp. 445–473.

7. ACKNOWLEDGEMENTS

The research presented here is part of the Nexus project which is supported as a Center of Excellence called "SPATIAL WORLD MODELS FOR MOBILE CONTEXT-AWARE APPLICATIONS" under grant SFB 627 by the Deutsche Forschungsgemeinschaft (DFG - German Research Council).

The test data have kindly been provided by the NavTech company (GDF) and the state survey office of the federal state of Baden-Wuerttemberg (ATKIS).