INTEGRATION OF GEOSCIENTIFIC DATA SETS AND THE GERMAN DIGITAL MAP USING A MATCHING APPROACH

G. v. Gösseln, M. Sester

Institute of Cartography and Geoinformatics, University of Hannover, Appelstr. 9a, 30167 Hannover, Germany – {guido.vongoesseln, monika.sester}@ikg.uni-hannover.de

Commission IV, WG IV/7

KEY WORDS: Cartography, GIS, Geology, Soil, Change Detection, Integration

ABSTRACT:

The integration of various data sets can be the answer for geoscientific questions on the one hand, but a disadvantage on the other hand, due to the differences in representation and content. Although geoscientific data sets typically refer to the same physical data source – the earth surface – and therefore also relate to topographic objects, these data sets differ in geometry, accuracy and actuality in most cases. In former times differences between analogue maps were not as apparent as today when different data sets are overlaid in a modern GIS-application. Integrating different data sets – in our case topographic data and geoscientific data – allows for a consistent representation and thus for the propagation of updates from one data set to the other. This problem leads to three steps, namely harmonisation, change detection and updating which are necessary to ensure consistency, but hardly practicable when performed manually.

For a harmonization of data sets of different origin, firstly the revelation of semantic differences is required; to this end, the object catalogues are compared and semantically corresponding objects are identified. In this step, also the cardinality of possible matchings between the objects in the different representations is determined (1:1, 1:n, n:m). The identification of geometric differences between the one-layered geoscientific and the multi-layered German digital map (ATKIS) will be fulfilled in the next step. In order to identify corresponding object-pairs between the data sets, different criteria like area, shape and position are used. Due to different levels of generalisation the detection of matches between groups of objects and single objects is implemented. Corresponding objects which have been selected through semantic and geometric integration are investigated for change detection using intersection methods.

The geometric differences which are visible as discrepancies in position, scale and size due to simple superimposition will lead to unsatisfying results. Therefore, the iterative closest point (ICP) algorithm is implemented to achieve the best fit between the objects. The evaluated results can be classified into three types, of which two types can be handled automatically, and for one type an automatic proposal is given by the software. This leads to a significant reduction of time and resources because the approach reduces the objects to be investigated manually to only those situations where manual intervention is inescapable.

The paper gives an overview of the problem and focuses on the geometric integration, especially on the matching of groups of objects and the adaptation of the object's shape.

1. INTRODUCTION

Geoscientific and environmental problems often require the usage of different data sources to achieve a satisfying result. The combination of different data sources offers the advantage to benefit from their respective merits. In former times these data sets were used in only analogue representations, but today the main part of geoscientific data sets are available as digital data sets.

The data sets which have been acquired for geoscientific purposes rely on the same source, the earth surface.

Despite this fact they show significant differences due to different acquisition methods, formats and thematic focus, different sensors, level of generalisation, and even different interpretation of a human operator. Sometimes new acquisition is therefore needed to create a single homogenous data set.

Another problem which occurs while working with different data sets is the problem of temporal inconsistency:

Even if the data sets originally are related to the same objects, different update cycles in the different thematic data sets lead to significant discrepancies. Observing this problem it is obvious that harmonisation, change detection and updating of different data sets is necessary to ensure consistency, but hardly practicable when performed manually.

Professionals from different geoscientific domains in Germany take advantage of the geological (GK) and the soil-science map (BK). These maps have a very strong thematic focus, but they do not contain the amount of topographic content, which is mandatory for different tasks to be solved. Therefore these data sets are combined with the german digital topographic data set (ATKIS). Unfortunately these data sets have been digitized from analogue maps and they differ in acquisition time, representation type and temporal consistency. which makes integration hardly possible.

In a project of the German Ministry of Education and Research under the headline "GEOTECHNOLOGIEN", a research group at the University of Hannover, consisting of three institutes from surveying and computer science, is dealing with the problem of data integration, applied to data sets from topography, geology and soil science. The project deals with different aspects of data integration, namely integration of different vector data sets, integration of vector and raster data, as well as providing an underlying data structure in terms of a federated data base, allowing a separate, autonomous storage of the data, however linked and integrated by adapted reconciliation functions for analysis and queries on the different data sets (Sester et al., 2003).

This paper focuses on the work of the Institute of Cartography and Geoinformatics (ikg), namely the integration of vector data.

Methods for the automatic identification of corresponding objects, adjusting the object geometry, and detection of changes which occurred in reality, but are not yet integrated in one of the data sets, will be developed. This is done with a focus on the above mentioned data set. Geometric aspects and methods will be described, namely the merging of segmented objects and the adaptation of the geometry by using a rigid transformation, followed by a mere intersection and evaluation of the resulting elements.
In this project the German digital topographic data set (ATKIS) will be chosen as reference, therefore the geometry of the geoscientific maps will be adapted without using constraints regarding accuracy or actuality so far. The approach, however, will be extended in the near future, to also take the relative accuracy and importance of the objects to be integrated into account.

2. RELATED WORK

Data can be integrated and fused for mutual benefit: Walter & Fritsch, (1999) present an approach that fuses two different data sets with road information with the aim of mutually exchanging attributes of the two data sets. The integration of vector data and raster data is being investigated in a GEOTECHNOLOGIEN partner project with the aim of enriching a 2D-vector data set with 3D-information (Butenuth & Heipke, 2003). Data integration or data matching is also needed for update purposes, e.g. when a data provider has to deliver up-to-date information details to his customers (Badard, 1999).
A conflation component strategy to provide independent but interoperable modules to solve special integration problems has been developed by Yuan & Tao, (1999).
Integration can be used for data registration, when one data set is spatially referenced and the other has to be aligned to it (Sester et al., 1998). A conceptual framework for the integration of geographic data sets, based on a domain ontology and surveying rules, was developed for update propagation between topographic data sets (Uitermark, 2001).
Finally, data integration is needed for the generation of Multiple Resolution Data Bases (MRDB); in this case objects of different geometric and thematic resolution have to be fused (Mantel, 2002).

3. USED DATA SETS

For the research in the GEOTECHNOLOGIEN project three data sets are used: the topographic data set ATKIS, the geological map and the soil-science map, all at a scale of 1:25000. When going from analogue to digital maps, new possibilities for data handling and analysis appear: basically, the combination of different data sets in a geo-information system (GIS) is enabled.
Simple superimposition of different data sets already reveals visible differences (Fig. 1). These differences can be explained by comparing the creation of the geological, the soil-science map and ATKIS (Goesseln & Sester, 2003).
As for ATKIS the topography is the main thematic focus, for the geo-scientific maps it is either geology or soil science, these maps have been produced using the result of geological drills and according to these punctual informations, areal objects have been derived using interpolation methods based on geoscientific models. However they are related to the underlying topography. The connection between the data sets has been achieved by copying the thematic information from topography to the geo-scientific maps at that point of time the geological or soil-science information is collected. This is done by using up scaled copies (1:25.000 to 1:10.000) of topographic maps. The selection and integration of objects from one data set to another one has been performed manual and in most of the cases the objects have been generalized by the geoscientist.
While the geological content of these data sets will keep its actuality for decades, the topographic information in these maps do not: In general, topographic updates are not integrated unless new geological information has to be inserted in these data sets.
The update period of the feature classes in ATKIS varies from one year up to three months – in general, 10% of the objects have to be updated per year (LGN 2003).

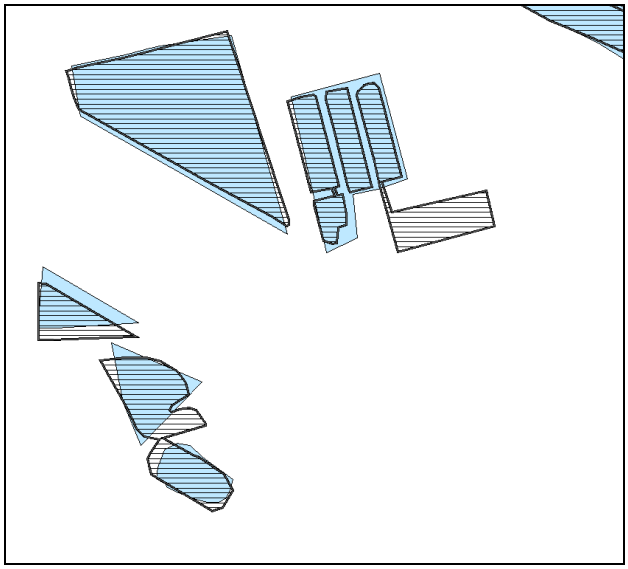

Fig. 1 : Simple superimposition of ATKIS (dark border, hatched) and geological map GK 25 (solid fill).

The geoscientific maps have been digitized to use the benefits of digital data sets, but due to the digitalization even more discrepancies occurred.
Another problem which amplifies the deviations of the geometry is the unequal data model between these data sets.
Geological and soil-science maps are single-layered data sets which consist only of polygons with attribute tables for the representation of thematic and topographic content, while ATKIS is a multi-layered data-structure with objects of all geometric types, namely points, lines and polygons, equally with attribute tables.
These differences in acquisition, creation, modelling and updating lead to discrepancies, making these data sets difficult to integrate. The amount of financial and human resources which is needed for the removal of these discrepancies can hardly be afforded. Therefore, new methods are required which offer an automatic or semi-automatic process capable of detecting and removing the differences between these data sets and supporting a human operator in this process.
In order to identify changes in the data sets and update the changes, the following steps are needed: identification of corresponding objects in the different data sets, classification of possible changes, and finally update of the changes.

4. DATA INTEGRATION

4.1 Overview

Data Integration is a very actual research topic covering many different aspects from a variety of different domains. In this part of the GEOTECHNOLOGIEN project the integration of heterogeneous vector data sets is the main focus. Data integration or map conflation can be divided in horizontal and vertical integration. *Horizontal conflation* is referred to edge-matching of adjacent maps with the objective of eliminating spatial and thematic discrepancies in the common area of the maps, *vertical conflation* describes the integration of two (or more) maps covering the same area with differences in data modelling, thematic content and accuracy (Yuan & Tao, 1999).
The result of the integration of ATKIS and the geoscientific maps is slightly different from the common definition of map conflation. As it is not the aim of the project to develop a new master data set (Beller et al., 1997), but to enhance the geometric accuracy of the geoscientific data sets. In this project the creation of a master set is not recommended because ATKIS is chosen as reference data set regarding the higher geometric accuracy and actuality. Therefore the topographic content of the geoscientific data sets is adjusted to a reference data set.
During the integration process there are various mandatory tasks. The geometric accuracy of ATKIS – which is based on the higher acquisition accuracy and the more frequent updates – should be used to correct and enhance the geometric content of the geoscientific data sets and avoid parallel updating.

4.2 Semantic Differences

At the beginning of the integration process the semantic models – which means at this time of the project the thematic contents – of all data sets are compared. Topographic elements which are represented in all of the three data sets are selected and will be used as candidates for the matching process. This selection is mandatory to avoid comparing "apples and oranges" and has to be the first step to ensure a successful integration.
Four different types of data integration are defined in (Walter & Fritsch, 1999).

- I.: stemming from the same data source with unequal updating periods,
- II.: represented in the same data model, but acquired by different operators,
- III.: stored in similar, but not identical data models,
- IV.: from heterogeneous sources which differ in data modelling, scale, thematic content.

The integrational part to be performed in this project could be categorized as type IV.
In the first phase of this project, the topographic feature class "water areas" has been chosen as a candidate for developing and testing, because of the presence of this topographic element in all data sets.

5. INTEGRATION WORKFLOW

One aim of the project is the adaptability of the research results to real applications. Therefore all the research is pursued in close partnership with external partners from geology and soil-science.

5.1 Application framework

At this point of the project the first research results and selected algorithms have been implemented in a software prototype.
Vividsolutions developed an open-source GIS application based on the JAVA development language. The Unified Mapping Platform JUMP is a GUI-based application for viewing and processing spatial data. It includes many spatial and GIS functions. It is also designed to be a highly extensible framework for developing and running customized spatial data processing applications. JUMP is based on the Java Topology Suite JTS, a JAVA programming library which offers various modules for the development of highly adopted software applications for data integration (JUMP 2004).
Using this system which represents data according to the OGC-standard a software prototype is developed, which serves as testbed for different matching-algorithms and is used for visualization of the origin data sets and the matching results.
The concept the federated database foresees that all the original data sets will be kept – however the links between corresponding objects in the different data sets will be explicitly stored.

5.2 Data preparation

Before the integration process can be started, all the data sets which will be used in the integration workflow, have to be pre-processed to a common data format.
In this project a federated data base is developed which is capable of importing the data sets in their original format, converting them to a common standard and store them in a single data management system (Tiedge et al. 2004).

5.2.1 Harmonisation

Water objects in ATKIS are represented in two different ways: Water areas and rivers exceeding a certain width are represented as polygons. Thinner rivers are digitised as lines and are assigned additional attributes, referring to some classified ranges of widths. The representation of water objects in the geo-scientific maps is always a polygon.
These differences have to be adjusted before integration starts. For the first implementation a simple buffer algorithm has been chosen, using the line representation from ATKIS as centre line and the width attribute. This enables the operator to compare the polygon from ATKIS and the water object from the geo-scientific maps using a mere intersection.
Another problem is the representation of grouped objects in different maps. For a group of water objects, e.g. a group of ponds, the representation in the different data sets could either be a group of objects with the same or a different number of objects, or even a single generalised object. Finally, also objects can be present in one data set and not represented in the other.
All these considerations lead to the following relation cardinalities that have to be integrated: 1:0, 1:1, 1:n, and n:m.

5.3 Geometry based matching

5.3.1 Selection Sets

As it was mentioned in 4.2 the data delivered from the data management system, will be selected using specified feature attributes, resulting in the three selection groups (ATKIS, geological map and soil-science map).
Due to the fact that the objects from all three data sets are representations of the same real world objects, they show

apparent resemblance in shape and position. The discrepancies between the data sets based on the different ways of acquisition, modelling and updating have been described at the beginning. But due to the diversity in digitizing the analogue geoscientific source maps and the data modelling of ATKIS, objects representing the same real-world objects differ in the number and geometry of segments (see Fig. 2)
Thus, investigating corresponding partners between the ATKIS and the geoscientific data sets, would lead not only to unsatisfying results but to relation errors. Therefore the investigation for corresponding objects has to be performed based on the aggregation of segments.
Using an overlapping test and by evaluating the overlap-area composed to the area of the segments to be tested, selection sets will be build, these selection sets will be stored as aggregated groups (with 1 to n elements). In order to find valid correspondences, all possible pairs of combination of neighbour objects will be checked against each other in the search process (see Fig. 3). Alternatively, we can use a breadth search procedure for finding the object clusters.

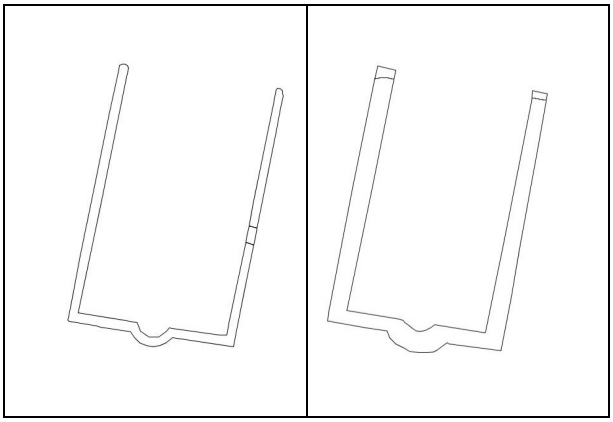

Fig. 2 : Segmented objects from the reference data set ATKIS (left image), and from the geological map (right image).

In order to define the neighborhood, either a buffer with a fixed distance or a triangulation can be used. A parameter free approach to identify clusters is based on an hierarchy of neighborhood graphs (Anders 2003).

5.3.2 Geometry based matching

The matching of the selection sets (e.g. the aggregated segments) will be checked individually using different measures.
In the current prototype the following measures for determining object similarity are used:

- **Hausdorff distance**: The length of the greatest local deviation between the two shapes. The lower the deviation, the higher the score.
- **Symmetric difference**: The areas found in one shape only. The more the two shapes overlap, the lower the symmetric difference, and the higher the score.
- **Compactness difference:** The difference between each shape's compactness, which is the area-to-perimeter ratio. The more similar the compactness of the two shapes, the higher the score.
- **Angle Histogramm:** The difference between each shape's angle histogram, which is a histogram of the angles that the segments make with the positive x-axis, weighted by segment length. The more similar the histograms for the two shapes, the higher the score.

For each geometric criterion a result between 0 and 1 is calculated and the mean value for each correspondence is evaluated. Different combinations of segments from the selection set of one data set are tested with the corresponding selection set (e.g the combinations of segments) from another data set . The highest result between to segment combinations will be kept as link. This process will be repeated until no more appropriate links can be established.

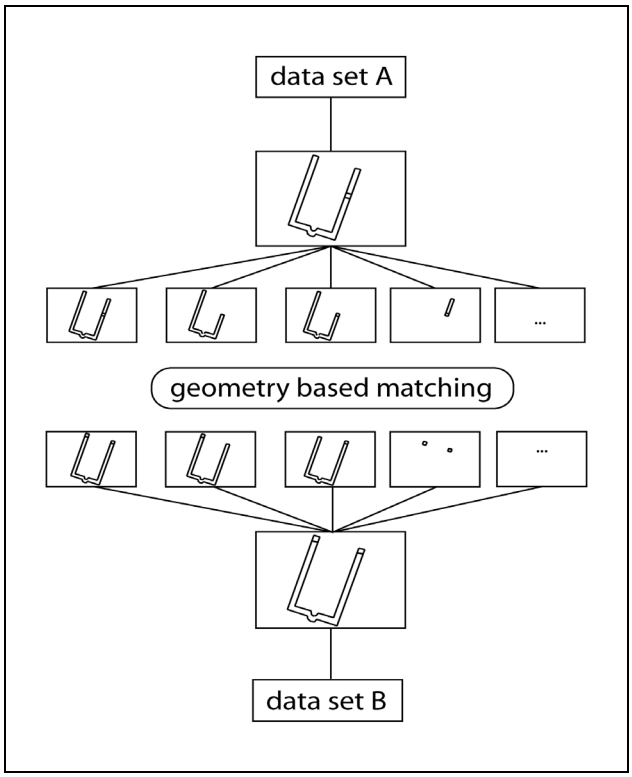

Fig. 3 : Selection set for geometry based matching between objects from two different data sets.

Once the correspondences between the selection sets have been found in the matching step, it has to be decided, whether the objects correspond exactly or if they differ due to update processes, which have been applied to one data set, but not to the other one. The automatic investigated links will be visualized to the operator, but before the next step – the change detection – will be performed, a manual correction of the links will be possible. Depending on geometric descrepancies, different types of change can be identified (see section 6.1).

6. CHANGE DETECTION

Objects which have been selected through geometric integration and have been considered as a matching pair could be investigated for change detection. A simple intersection of corresponding objects is used for the change detection. Yet, the mentioned differences may cause even more problems which are visible as discrepancies in position, scale and shape. These discrepancies will lead to unsatisfying results and make the evaluation of the resulting elements almost impossible (Fig. 4).
Therefore firstly, a local transformation will be applied, leading to a better geometric correspondence of the objects. To this end, the iterative closest point algorithm (ICP) developed by (Besl &

McKay, 1992) has been implemented to achieve the best fitting between the objects from ATKIS and the geo-scientific elements using a rigid transformation.
In our first approach, objects from ATKIS are considered as reference due to their higher geometric accuracy, and the objects from the geoscientific datasets are optimally fitted to the ATKIS objects (Goesseln & Sester, 2003).

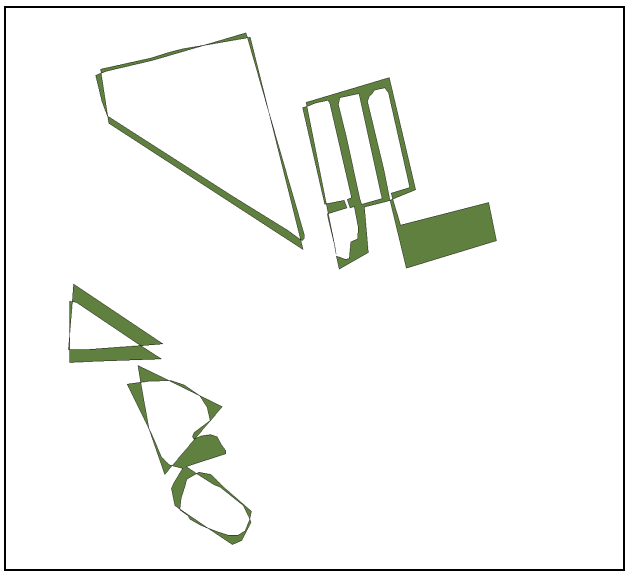

Fig. 4 : Resulting overlapping segments from mere intersection showing geometric differences between water bodies in the German digital topographic map (ATKIS) and in the geological map.

At the end of the process the best fit between the objects using the given transformation is achieved, and a link between corresponding objects in the different data set is established. The ICP algorithm has been implemented to compensate the geometric discrepancies which occur due to the way the digital geoscientific data sets have been created using manual adaptation, rescaling and digitization.

6.1 Intersection and segment evaluation

Following these steps, intersecting objects for a proper change detection will lead into a more reliable result (Fig. 5) than simple intersection (Fig. 4). This analysis and the classification into different change situations is a semantic problem and will be conducted in close collaboration with experts from geology and soil science, who are also partners in the project.
At this time of the project three different classes have been identified: the intersection segments can be classified according to their respective classifications in the original data sets in:

- Type I : Segment is defined as water area in both maps, no adaptation required,
- Type II : Segment in geoscientific data set has been any type of soil, but is defined as water-area in the reference data set; therefore the attribute of classification will be changed in the geoscientific map,
- Type III : Segment is defined as water-area in geoscientific data set (e.g. no soil-type definition available), but no water-area in the reference data set. Therefore a new soil-definition is required.

Type II will also be assigned to objects which are represented in the reference, but not the candidate data-set, this is the result different updating periods between the reference and the candidate data set, which results in outdated objects.
While Type I and II require only geometric corrections or attribute adaptation and can be handled automatically, Type III needs more of the operators attention.
Depending on the size and the shape of a Type III segment and by using a user-defined threshold, these segments can be filtered, removed and the remaining gap can be corrected automatically, this will avoid the integration of sliver polygons and segments which are only the results of geometric discrepancies and must not be taken into account.
Different situations can cause the presence of a Type III segment. Due to different natural effects like desiccation or man-made rerouting of a river bed, water areas have been changed in shape or they even disappeared from the face of the environment.
After an actual topographic description is no longer available, there is no up to date process or method to derive a new soil definition automatically. As there are different ways an water area can disappear, there are different natural (e.g. erosion) or man-made (e.g. refill) processes which have influence to the new soil type. This new soil type could not be derived automatically, but there are different proposals which could be offered to the user by the software. An area-threshold which has to be defined in the near future together with the experts from geology and soil-science will be applied to remove Type III segments which occur due to geometric discrepancies.
As a result a visualisation will be produced showing all the areas where an automatically evaluation of the soil situation could not be derived or only a proposal could be delivered and manual “field work” must be performed (Fig. 5).
The visualisation of Type III segments will already reduce the amount of human resources needed to detect the topographic changes between the geoscientific data sets and ATKIS.
It is expected, that a high degree of automation can be achieved with this process. In some situations there will be an automatically generated suggestion from the algorithm, however the expertise of a human operator will still be mandatory in some cases in order to commit or propose another solution.

7. CONCLUSION AND OUTLOOK

The workflow presented in this paper is the result of the research and has been developed in close correspondence with the project-partners from geology and soil-science.
The implementation of the workflow in a software protoype using the open source software JUMP will ensure the possibility of adopting the results of this project to any additional vector-vector integration.
The implementation of the filtering, geometric comparison and the derivation of object links, together with the ICP-algorithm showed very good results. Processing the test data set, representing a standard geoscientific data sets needs less than a minute for water-areas.
At this point of the project one data set is selected as reference data set, which will remain unchanged while the candidate data sets are adjusted. If an even more accurate correspondence between the data sets is needed, specific geometric reconciliation functions for the exact adaptation of the geometry have to be implemented. The idea is that for that purpose, the individual shapes of the objects will be geometrically adjusted: depending on the relative accuracies of the original objects, an “intermediate” geometry will be

calculated. This will be achieved using a least squares adjustment process, where observations in terms of differences in shape will be introduced as a functional model – the stochastic model will describe the accuracies of the original shapes. This process then will lead to a local adaptation of the individual corresponding objects, but also of their local environment. Too large discrepancies of the shape boundaries will be considered as outliers and can be treated in the subsequent overlay and analysis step.

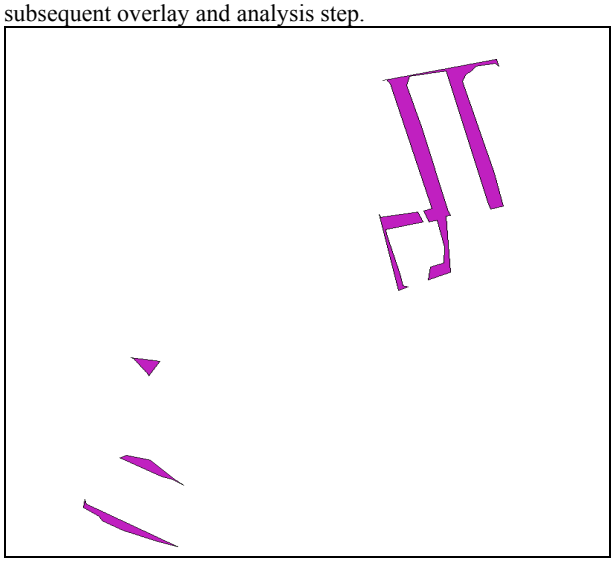

Fig. 5 : Visualisation of changes between topographic content from ATKIS and geological map, after applying ICP algorithm and area-threshold filtering.

At the end of the project constraints for every data set can be defined which will facilitate the creation of a weighted geometry or a so called master data set which is the common idea of map conflation.
In the near future the introduction of punctual and linear elements will enhance the process of geometric integration, because at this stage of the project only polygons are evaluated.
Further work will concentrate on partial matching that often occur at object boundaries: e.g. a geoscientific object ends at a river or road. This means, that these features have a part of the river or road boundary in common. In order to identify these partial correspondences, it is necessary to appropriately segment these objects.
Due to the fact that only the geometry of linked objects is changed and adjusted during the workflow the neighborhood remains unchanged. These discrepancies will be removed at the end of the integration process, to ensure a topologically consistent model, the data management system from the federated data base is capable of validating the topology structure to avoid saving corrupt data.

ACKNOWLEDGEMENTS

This is publication no. GEOTECH-66 of the GEOTECHNOLOGIEN project funded by the Federal Ministry for Education and Research (BMBF) and the German Research Council (DFG) under contract 03F0374A.

REFERENCES

Anders, C.: A Hierarchical Graph-Clustering Approach to find Groups of Objects, (Technical Paper), *ICA Commission on Map Generalization, Fifth Workshop on Progress in Automated Map Generalization,* IGN, Paris, 28-30.04.03.

Badard, T., 1999. On the automatic retrieval of updates in geographic databases based on geographic data matching tools. In: *Proceedings of the 9th International Cartographic Conference*, Ottawa, ICA/ACI (Eds.), 1999, pp. 47-56.

Beller, A., Doytsher, Y., & Shimbursky, E., 1997. Practical Linear Conflation in an Innovative Software Environment. In: *1997 ACSM/ASPRS Annual Convention and Exposition Technical Papers*, Seattle, Washington, April 1997, vol. 2, pp. 146-153.

Besl, P. & McKay, N.,1992. *A Method for Registration of 3-D Shapes,* Trans. PAMI, Vol. 14(2), pp. 239-256.

Butenuth, M. & Heipke, C., 2003. Modelling the integration of hetergeneous vector data and aerial imagery. In: *Proceedings of ISPRS Commission IV Joint Workshop*, Stuttgart, Germany, September 8-9, 2003. Submitted.

Goesseln, G. v. & Sester, M., 2003. Semantic and geometric integration of geoscientific data sets with atkis – applied to geo-objects from geology and soil science. In: *Proceedings of ISPRS Commission IV Joint Workshop*, Stuttgart, Germany, September 8-9, 2003.

JUMP, 2004. The unified mapping platform, http://www.vividsolutions.com, (visited May 2004).

LGN, 2003. ATKIS in Niedersachsen und in Deutschland. In: *Materialien zur Fortbildungsveranstaltung Nr. 1/2003*, Hannover.

Mantel, D., 2002. *Konzeption eines Föderierungsdienstes für geographische Datenbanken.* Diploma thesis, unpublished, University of Hannover.

Sester, M., Hild, H. & Fritsch, D., 1998. Definition of Ground-Control Features for Image Registration using GIS-Data. In: *T. Schenk & A. Habib, eds, 'IAPRS', Vol. 32/3, ISPRS Commission III Symposium on Object Recognition and Scene Classification from Multispectral and Multisensor Pixels*, Columbus/Ohio, USA, pp. 537-543.

Sester, M., Butenuth, M., Goesseln, G. v., Heipke, C., Klopp, S., Lipeck, U., Mantel, D., 2003. New methods for semantic and geometric integration of geoscientific data sets with ATKIS – applied to geo-objects from geology and soil science. In: *Geotechnologien Science Report, Part 2*, Koordinierungsbüro Geotechnologien, Potsdam.

Tiedge, M., Lipeck, U. & Mantel, D., 2004. Design of a Database System for Linking Geoscientific Data. In: *Geotechnologien Science Report, Part 4*, Koordinierungsbüro Geotechnologien, Potsdam.

Uitermark, H., 2001. *Ontology-based geographic data set integration*. Doctor thesis, Deventer, Netherlands.

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

Yuan, S. & Tao, C. 1999. Development of conflation components. In: Li, B. et al., eds., Geoinformatics and Socioinformatics, The proceedings of Geoinformatics´99 conference, Ann Arbor, pp. 1-13.