Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

An Environment for the Conceptual Harmonisation of Geospatial Schemas and Data

Thorsten Reitz
Fraunhofer-Institute for Computer Graphics Research (IGD)
Fraunhoferstraße 5
Darmstadt, Germany
thorsten.reitz@igd.fraunhofer.de

Simon Templer
Fraunhofer-Institute for Computer Graphics Research (IGD)
Fraunhoferstraße 5
Darmstadt, Germany
simon.templer@igd.fraunhofer.de

Abstract

We present a system for defining and evaluating conceptual schema mappings for the harmonisation of geospatial data sets. The goal of this system is to allow domain experts to ensure logically and semantically consistent mappings and consequently high-quality transformed geospatial data. Furthermore, a major focus is put on documentation of the schema transformation process and its impacts, e.g. in the form of lineage information attached to the resultant transformed data. This paper provides a summary of our mapping methodology, outlines lessons learnt in the first two evaluations and explains how the architecture we implemented evolves to meet upcoming user requirements.
Keywords: Schema mapping, Data Harmonisation, Systems and Architectures, User interface technologies, Geospatial data, Spatial Data Infrastructures

1 Introduction

Spatial Data Infrastructures (SDIs) are becoming more and more widespread, as indicated by national and international reports [1, 2]. Data providers offer spatial data sets through standardized service interfaces and make use of nationally and internationally specified data models, such as ATKIS in Germany or the INSPIRE Data specifications for Europe. However, the improved provision of spatial data services does not solve all integration problems faced by data users, especially when data from different providers and with different lineage is to be used together. Even data delivered over the same service interface, in the same conceptual model can be very heterogeneous in its actual content. Typical heterogeneity issues that make usage of integrated data hard include different handling of multiple representations, different classification systems and different interpretations of standards. Full interoperability consequently requires data harmonisation, which we define as "*creating the possibility to combine data from heterogeneous sources into integrated, consistent and unambiguous information products, in a way that is of no concern to the end-user*"[3].

SDIs as they are implemented today reduce the integration burden for clients, but they do not yet take away the *harmonisation* burden from these clients. Supporting spatial data users in data harmonisation is a crucial aspect of making SDIs successful. Otherwise, harmonisation aspects have to be handled by every client on their own – and harmonisation issues are complex and hard to resolve. As a result, SDI resources are not used as well as they could be for processes such as reporting, data fusion and analysis. We therefore work on approaches for data harmonisation services to augment SDIs.

One specific aspect is the semantic harmonisation of the geospatial data, for which two essential steps have to be completed. First, a harmonised conceptual schema needs to be designed, and second, conceptual schema mappings have to be created [4]. A schema mapping is a formal description of how concepts from one conceptual schema relate to concepts from another conceptual schema. Such a formal description consists of statements on the nature of the relationship of the concepts and can be used to derive concrete rules how instances of one concept can be transformed into instances of the other schema's concepts. Creating these mappings can be a very complex task that requires both domain expertise and expertise in formal modelling approaches.

As of today, a common approach is that a domain expert fills out specification spreadsheets, so called matching tables, and passes these on to IT experts. The IT experts then implement the required transformations, often on the basis of specialised Extract/Transform/Load (ETL) software such as FME or on the basis of generic purpose tools, such as XSLT [5]. This approach leads to relatively long iterations, with a duration of a few hours to many months. However, efficient schema and data harmonisation activities require much faster feedback times. As an example, as a verification step in the Data Specification work performed as part of the INSPIRE implementation, transformation testing (i.e. the mapping existing data to harmonised draft schemas) was a mandatory activity that took months to complete, and was done on a highly heterogeneous technical basis.

Another open issue with current mapping and transformation approaches is the quality assurance of the mapping. Especially the identification of mismatches, i.e. subtle structural and semantic differences in concepts that are mapped, is not supported well. This support can be provided by applying automated identification methods and visualization approaches typical for geospatial data to relate the effects of mappings and to see possible mismatches. Therefore, such approaches and tools can assist domain experts in creating high-quality mappings and ensuring interoperability of schemas as an enabler of SDIs.

1.1 Contribution and Structure

The main focus of this work is to provide methods for the integration of two or more existing (geospatial) data sets into a common conceptual schema, with consistently high and documented quality, and to implement and evaluate these methods. The system presented in this paper takes up the declarative and expressive ontology mapping approach (EDOAL) introduced by David, Scharffe et al. [6]. Instead of relying strictly on formal ontologies, which are very rarely available for spatial data sets, we use EDOAL with different types of schemas – from UML models to XML Schema Definitions. To offset the disadvantages compared to a full ontology that these schema types bring, we furthermore use automated analysis methods on the schema and on instances where available for the quality management of the mapping. This quality management comprises two components – documentation of mapping limitations [7] and debugging mappings. The two main aspects that form the contribution of this paper are the evaluation of our system for the interactive and declarative mapping for geospatial domain experts as it was at the end of the HUMBOLDT project, and the description of the evolution of this system based on the findings of this evaluation.

This paper is structured as follows: in the following section we explore which approaches for conceptual schema mapping are available and which are the specific shortcomings of the individual approaches with respect to the application field of geospatial schema/data integration. The next section sketches the interactive and declarative conceptual schema mapping approach we are implementing and provides a detailed view of the implementation. In the final two sections, evaluation results, a conclusion and the outlook are presented.

2 Mapping Approaches and their Applicability in Geospatial Schema Mapping

There are several different types of approaches for the conceptual and physical integration of geospatial schemas and data sets. From a user's perspective, the core goal is the integration and harmonisation of heterogeneous data sets, and not only the reconciliation of conceptual models such as ontologies.

A full integration of geospatial data therefore requires both schema mapping and data transformation. The schema mapping is done to find conceptual relationships and inconsistencies while data transformation deals with the particularities of handling different attribute types and structures. In the sense of a model-driven architecture, the schema mapping is defined at the level of the conceptual schema, whereas the transformations are defined on the level of the logical schema. The actual execution of the transformation then takes place at the instance level.

There are several different types of approaches for the conceptual and physical integration of geospatial schemas and data sets. Most existing approaches tend to focus on one level and do not address the whole stack, from the purely conceptual model down to the physical encoding of the instances:

- *Intension-centric* integration approaches focus on defining the relationships between the elements of the conceptual schema of the data to be integrated [8-10];
- *Extension-centric* integration approaches focus on the transformation process definition to translate instances from a source schema to a target schema [11-13];

There are also integration approaches that take into account both the conceptual schema and the instance information, such as the ORCHID [14] and SPICY [15] systems, as well as the "*Data-Driven Matching of Geospatial Schemas*" approach described by Volz [16].These approaches focus on exploiting instance relationships for improving the automated matching on the schema level, but do not provide instance transformation derived from the conceptual schema mapping.

An issue that reduces acceptance of the intension-centric approaches is that they require the specification of a formal conceptual schema, e.g. using an Ontology language such as OWL. Such ontologies or even simple taxonomies are not created or used by geospatial data experts on a regular base, despite the big number of top-level and application-level ontologies that have been created by the research community [17]. As an example, for the more than 300 data sets that were harmonised as part of the HUMBOLDT project, only a single one had a corresponding ontology, and for that one, no vertical mapping had been defined to link to the logical schema of the data. This lack of a formal semantic definition means that users perceive a hurdle to using semantic web technologies, despite all the potential advantages they have, such as ensuring formally consistent mapped conceptual schemas [18].

A second issue is user trust in the matching methods that are being developed. In general, only manual and semi-automated processes were accepted by the users from the geospatial domain that participated in the specification of our approach. Furthermore, automated matching doesn't necessarily bring the same level of benefit to the geospatial field as to other application domains: In geospatial schema mapping, schemas are often very complex along the property structure, but are not as complex in the inheritance structure of the classes. Compared to medical or biological ontologies, the number of classes is low; at the same time, aggregation structures in geospatial classes can be five or more levels deep.

Approaches that focus only on the extension and its transformation have higher acceptance rates; especially ETL-type systems are common. However, it requires high effort to define the mapping between increasingly complex schemas, and resulting mappings can get complicated and unmaintainable [19]. To reduce effort and complexity, a current approach is to provide model-specific mapping definition UIs, e.g. for the INSPIRE specifications[1]. In a world where new data sets and models are created all the time or existing ones change, this approach is not sustainable.

Also, information that can be used to assess the quality of the mapping which is available in a conceptual model is not accessible to extension-centric ETL approaches.

[1] Examples include the conterra FME INSPIRE Solution Pack or the AED SICAD FUSION Data Service.

3 A System for Interactive and Declarative Schema Mapping

A declarative approach for schema mapping has the advantage that the created mappings are of minimal complexity; the amount of inputs that needs to be made is much lower than with procedural approaches. The interactive approach ensures understanding and transparency to the user. Both are essential, as the main purpose of our approach is to assist a domain expert in creating a schema mapping that will enable him to transform data conforming to a source schema to a harmonized target schema – and to understand in detail what limitations the created mapping has and how these limitations impact the fitness-for-purpose of the transformed dataset.

The model for our declarative schema mapping approach is based on the Ontology Mapping Language (OML), which has now evolved into the Expressive Alignment Format [6]. A schema mapping defines relations between source and target entities, which represent the source and target schema elements. For a mapping between conceptual schemas these entities represent classes and properties. Logical schemas can be abstracted to conceptual schemas, assumed they can be mapped vertically to the concepts of classes and properties, allowing the definition of the mapping on the conceptual level. The actual transformation then must again take the logical schema into account.

Due to the declarative form of the mapping definitions multiple mappings can be easily combined, allowing the definition of multiple transitive mappings resulting in a single transformation. This can serve for maintaining mappings when dealing with new versions or changes to the schemas. Also, defining a mapping between two schemas can thus be performed by creating and combining several partial mappings, e.g. each authored by the data expert most familiar with the corresponding entities. For schemas that have an extensive hierarchy of classes, the mapping effort can be substantially reduced by defining relations on super classes, which are propagated to their descendants, preventing the need to define these relations for each of the sub classes.

There are two basic types of relations; relations between classes and relations between properties. Property relations are only applicable in the context of a class relation. As such the task of defining a mapping can be split into several steps, first defining the relations between classes, then, in the context of each class relation, specifying the corresponding property relations. This allows a more focused view on the schemas when specifying the mapping, only considering information that is relevant in the context of the currently examined class relation.

Apart from single classes or properties an entity may also represent a composition of multiple classes or properties, e.g. where multiple property values in the source schema must be combined to yield the desired property value for the target schema. When dealing with cases where a source and target entity only represent a partial match, it is necessary to specify a domain of validity, e.g. limiting a relation to classes with certain property values. As information required for a target entity may only be available implicitly, a special type of relation is needed that allows assignments of information on target entities. In general, the actual relation between two entities can be of any kind, so there is the need to be able to define a huge variety of relations and the corresponding transformation functions, with possibly implementations for various transformation engines.

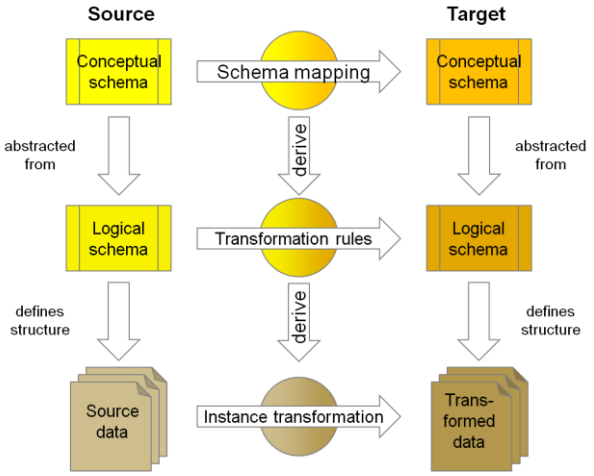
Our initial approach for the implementation of these functions was for them to have full control and responsibility on how to retrieve the source values, transform them and integrate the result into the target structure. This didn't proof feasible in the context of mappings targeting complex schemas, as the implementation effort for functions to support this is very high, and knowledge of other relations is needed to create a consistent target structure. As such we reduced the functions to provide only the value transformation, keeping them as simple as possible and moving the higher transformation logic to the transformer with access to the whole set of relations.

For the actual transformation there are some aspects that are not or only implicitly covered by the conceptual schema mapping:

- How to handle multiple occurrences of a property
- Data type conversion for property values
- Encoding of the target data

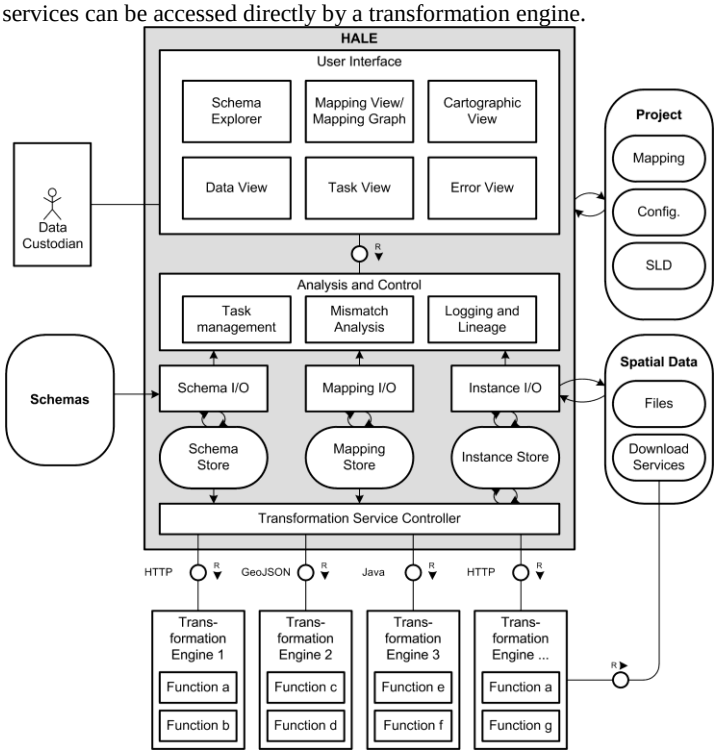
Taking into account the information from the logical schemas, transformation rules are derived from the schema mapping. These rules are then evaluated based on each concrete instance to be transformed to perform the transformation on the actual data (see figure 1). Automatically determining, how multiple occurrences of a property have to be represented in the target instance is only done in this last step, as this decision may be affected by multiple relations and the source instance structure. Also, validity constraints on relations can be evaluated only on the concrete instances.

Figure 1: Mapping and transformation levels



The interactive schema mapping approach is all about the feedback to the mapping author. Samples of the source data are used to give instant feedback on the consequences of changes to the mapping. During the process of defining the mapping, the resulting transformation is applied with each change, yielding transformed target instances, which can serve as measures for the current state and completeness of the mapping. This enables mapping authors to verify the transformation against their expectations of what the result should look like.

Figure 2: FMC Block diagram showing the core components of HALE and invoked Transformation and Data Services. Since HALE often works with subsets of larger spatial data sets, full sets offered by download services can be accessed directly by a transformation engine.



To this purpose, support for the target-oriented analysis of the data is required, e.g. the comparison of certain source instances with their transformed counterparts. As we are dealing with Geospatial data, much insight can be gained from the visual analysis of the data, displayed in a map – especially if the mapping involves transformation of geometries. For a thorough visual analysis the ability to define custom map styles has proven useful, e.g. by defining a style based on property values, thus identifying certain instances.

Additionally to the data itself, there is information on the mapping that can be derived from the conceptual and logical schemas. The structure and constraints on the schemas can serve to identify possible mismatches. Also, this information can be used to guide in the process of defining the mapping, checking if the constraints of the target schema are met.

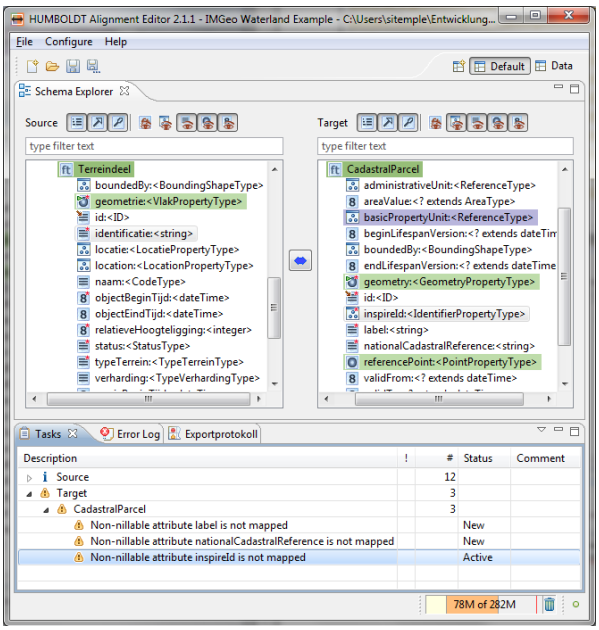
The general workflow for defining a mapping is as follows:

1. Load the source schema
2. Load the target schema
3. (optional) Load sample source data
4. Identify and create class relations
5. For each class relation identify and create property relations
6. (optional) Save the transformed data

For steps 4 and 5 with each change of the mapping the user has the possibility to analyze the current transformation result. Based on the current mapping tasks are generated for the user, that assist in choosing the next steps, e.g. pointing out mandatory target properties that need a relation to be defined.

The mismatch analysis and the task management system are used to debug and improve schema mappings, and to document known limitations. Towards this latter goal an in-depth analysis of mappings and available instance data is performed to find inconsistencies in the mappings as well as irreconcilable mismatches.

Figure 3: HALE project with open tasks (bottom) left for the mapping between the source (top left) and target (top right) schemas



This approach was implemented in the HUMBOLDT Alignment Editor (HALE). HALE provides an infrastructure

for schema mapping that can deal with the most common harmonization problems but can be easily extended to suit the needs of additional scenarios. The core components of HALE are pictured in figure 3. Currently there are several extensions available providing support for the following formats:

- Shapefiles, XML Schemas and GML application schemas as schema formats; experimental support for UML (in XMI or Enterprise Architect encoding) and OWL is also provided
- Shapefiles, XML and GML as instance formats
- OML, RIF-PRD and CSV as mapping formats
- HTML and CSV as documentation formats

4 Evaluation

The HUMBOLDT Alignment Editor has been undergoing different kinds of evaluation and validation to verify its utility and to improve the approach and implementation.
However, since the main goal of its inception was to support geospatial domain experts in creating high-quality mappings, the most important evaluation is that by the end users in a controlled test. Such a task-based usability test was conducted in May to September 2010 (based on version 2.0 of HALE) with the support of 30 users from the HUMBOLDT project, from Germany, Portugal, France, Italy and Hungary. The test consisted of a complex mapping task, relating a national cadastral parcels schema to the corresponding harmonised schema from INSPIRE. These users did not have prior knowledge of HALE and were classified as "Data Custodians", i.e. persons responsible for the maintenance of a certain spatial data set, with detailed knowledge of these data sets and their specifics. After conducting the test, the users were asked to answer a questionnaire, and the mapping was analysed for its completeness and correctness. The following table provides summary scores for selected sections of the questionnaires as well as summary scores for the mapping quality, which were calculated according to the mapping quality model presented in [19]). A similar study was conducted in 2009 for evaluating two different approaches – procedural programming with XSLT in XMLSpy and graphical pipes-and-filters definition using FME [19]. The results are included in Table 1 where they are comparable. This study served to define benchmarks for our approach.

Table 1: Quantitative findings of the evaluation
(scale: worst score is 0, best score is 1, values in parentheses are from 2009 precursor study)

Aspect	XSLT	FME	HALE
1.1 Mapping Correctness	0.80	0.90	0.94
1.2 Mapping Completeness	0.71	0.80	0.82
1.3 Mapping Speed	0.04	0.29	1.00
2.1.1 Schema import & handling	(0.80)	(0.71)	0.92
2.1.2 Data import & handling	(0.80)	(0.91)	0.67
2.2 Structure of the User Interface	(0.44)	(0.80)	0.79
2.3 Quality and utility of the Documentation	(0.61)	(0.87)	0.65
3.1 Semi-automated mapping support	-	-	0.68
3.2 Reclassification support	-	-	0.82
3.3 Merging of Features	-	-	0.40
3.4 Transformation functions	0.31	0.94	0.56
3.5 Mismatch understanding	0.40	0.55	0.80

In addition to these quantitative findings, the participants in the evaluation provided a lot of comments and ideas about how to improve the software and the approach. As an example, the hierarchical display of the inheritance hierarchy of the types proved to be of little value to most users, so the default presentation was changed in HALE. The full findings of the evaluation, together with tasks, questionnaires and other aspects of the methodology, have been documented in the HUMBOLDT project deliverable A10.3-D2[2].

Further feedback on HALE continually comes from the user community. The HUMBOLDT Alignment Editor is available as Free and Open Source Software[3] and has been downloaded about 4.800 times (as of 01/2012). It has been used in several projects now, ranging from training courses to production environments.

5 Conclusion & Outlook

In this paper, we describe an approach for the conceptual harmonisation of spatial data. We outline an interactive and declarative approach as well as the basic mapping process, followed by an overview of our proposed system architecture and its implementation HALE. This implementation was evaluated, showing high acceptance levels for the implemented workflow and most of the features HALE offers. It was also important to see that high-quality mappings are created, and that users understand limitations of these mappings better than with the approaches they have used before.

Our studies however also suggest incorporating the following improvements:

- Support for Linked Open Data structures;
- Better support of high cardinality mappings in complex structures, e.g. in geographic name structures. This is

[2] This report is available from http://www.esdi-humboldt.eu.
[3] HALE can be downloaded from http://community.esdi-humboldt.eu.

an aspect not managed well by any ETL or schema transformation software as far as we know, and the need for this was confirmed by the relatively low evaluation score for HALE's merging support;
- Handling and regeneration of references between instances;
- Improved collaboration and task management functionality;
- Direct editing and profiling of schemas would lead to improved user productivity, especially in transformation testing and target schema design activities.

6 Acknowledgements

We would like to thank Dominique Laurent from IGN France, who coordinated the evaluation of the HALE software, very much. HALE development was supported by Anna Pitaev, Andreas Burchert, Patrick Lieb and Sebastian Reinhardt. This work was partially funded under the HUMBOLDT Project, EC contract SIP5-CT-2006-030962.

References

[1] Janssen, K., Vandenbroucke, D.: Spatial Data Infrastructures in Europe: State of play 2006, *http://www.ec-gis.org/inspire/reports/stateofplay2006/INSPIRE-SoP-2006%20v4.2.pdf*, (2006).

[2] Vries, W.T.D., Crompvoets, J., Stoter, J., Berghe, I.V.: ATLAS of INSPIRE – conceptualizing SDI implementation through an inventory of experiences, successes and headaches of European national mapping agencies. *IJSDIR*. 6, (2011).

[3] Villa, P., Reitz, T., Gomarasca, M.: The HUMBOLDT project for data harmonisation in the framework of GMES and ESDI: Introduction and early achievements. *International Society for Photogrammetry and Remote Sensing - Proceedings of Commission IV*. S. 1741 - 1746. , Beijing, China (2008).

[4] Waters, R., Beare, M., Walker, R., Millot, M.: Schema Transformation for INSPIRE. *IJSDIR*. 6, (2011).

[5] Letho, L., Sarjakovski, T.: Schema Translations by XSLT for GML-Encoded Geospatial Data in Heterogeneous Web-Service Environment. *XXth ISPRS Congress*, Instanbul (2004).

[6] David, J., Euzenat, J., Scharffe, F., Santos, C.T. dos: The Alignment API 4.0. *Semantic Web*. 2, 3-10 (2011).

[7] Reitz, T.: A Mismatch Description Language for Conceptual Schema Mapping and its Cartographic Representation. *Proceedings of the 6th GIScience Conference (to appear).* , Zürich (2010).

[8] Nguyen, N.T.: A METHOD FOR ONTOLOGY CONFLICT RESOLUTION AND INTEGRATION ON RELATION LEVEL. *Cybernetics & Systems*. 38, 781-797 (2007).

[9] Cruz, I.F., Sunna, W., Makar, N., Bathala, S.: A visual tool for ontology alignment to enable geospatial interoperability. *J. Vis. Lang. Comput*. 18, 230-254 (2007).

[10] Agarwal, P., Huang, Y., Dimitrova, V.: Formal Approach to Reconciliation of Individual Ontologies for Personalisation of Geospatial Semantic Web. *GeoSpatial Semantics*. S. 195-210 (2005).

[11] Wang, S., Englebienne, G., Schlobach, S.: Learning Concept Mappings from Instance Similarity. In: Sheth, A., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., und Thirunarayan, K. (hrsg.) The Semantic Web - *ISWC 2008*. S. 339-355. Springer Berlin Heidelberg, Berlin, Heidelberg (2008).

[12] Letho, L.: Schema Translation in a Web Service Based SDI. *Proceedings of the 10th AGILE International Conference on Geographic Information Science*. , Aalborg, Denmark (2007).

[13] Frantz, R., Corchuelo, R., Gonzalez, J.: Advances in a DSL for Application Integration, (2008).

[14] Orchid: Integrating Schema Mapping and ETL. 1307-1316 (2008).

[15] Bonifati, A., Mecca, G., Pappalardo, A., Raunich, S., Summa, G.: Schema mapping verification: the spicy way. *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. S. 85–96. ACM, New York, NY, USA (2008).

[16] Volz, S.: Data-Driven Matching of Geospatial Schemas. *Spatial Information Theory*. S. 115-132 (2005).

[17] Buccella, A., Cechich, A., Fillottrani, P.: Ontology-driven geographic information integration: A survey of current approaches. *Computers & Geosciences*. 35, 710-723 (2009).

[18] Schade, S.: Computer-Tractable Translation of Geospatial Data. *IJSDIR*. 5, (2010).

[19] Reitz, T., Kuijper, A.: Applying Instance Visualisation and Conceptual Schema Mapping for Geodata Harmonisation. *Advances in GIScience*. S. 173-194 (2009)