Schema Transformation as a Tool for Data Reuse in Web Service Environment

Lassi Lehto
Department of Geoinformatics and Cartography
Finnish Geodetic Institute
Masala, Finland
lassi.lehto@fgi.fi

***Abstract*—The requirement to support the increased demand for geospatial data resources in several application fields, on various technical platforms and in diverse cultural contexts creates a big challenge for data providers. Reusing the same data set in different environments by employing content transformation processes is proposed as a solution. In this paper online schema transformation is evaluated as a tool for supporting data reuse in Web Service-based data delivery architecture. The case implementation is developed in the context of a major EU-project helping the National Mapping and Cadastral Agencies to achieve INSPIRE-compliance.**

Keywords-schema transformation; data reuse; Web processing; Web Services; ESDI

I. INTRODUCTION

A. *Background and motivation*

One of the most important benefits gained by the introduction of Web Services as the means for delivering digital geographic information is the increased use of the data. The extensive national geodata resources can potentially support many different applications and be used in various technical and cultural contexts, both on national level and abroad. Consequently, data providers are faced with increased need for producing ad hoc extracts from their databases and support a number of individual user preferences in the process. The diverse requirements concerning data encodings, schemas, coordinate reference systems and spatial representations are becoming a great challenge for data providers.

The number of use cases for network-accessible digital geodata has multiplied and the data sets are finding usage in many exotic applications and at the same time are in the core of many vital processes of society [1]. Significant additional requirements are set by the need for international co-operation and the subsequent need for data harmonisation, for instance in the context of the European cooperation [2].

Seamless provision of basic geospatial information is a necessity in the increasingly integrated Europe. Especially the challenges related to the protection of environment emphasise the need for providing access to geospatial information consistently across the European borders. The long history of national independence in Europe has contributed to the fact that each Member State (MS) of the European Union (EU) applies rather individual approaches as it regards collection, organisation and maintenance of geographic information. The challenge of integrating these heterogeneous sources of information as a reliable and consistent information service has been tackled in various Pan-European projects and initiatives, aimed at supporting the development of the European Spatial Data Infrastructure (ESDI).

The ESDI is currently under very active development. Various research programmes of the European Commission (EC) have included actions aimed at building components for this infrastructure. One of the most important drivers in this development is the INSPIRE (Infrastructure for Spatial Information in Europe) initiative [3]. The INSPIRE process is initiated by the EC as a European Directive and aims at seamless Pan-European spatial data provision in support of the protection of the environment [4].

A fundamental principle of the INSPIRE process is to build the European level geospatial content services on top of the existing National SDIs (NSDI), without explicitly requiring changes on the Member State data sets. The diversity of the solutions adopted by the European NSDIs makes this approach particularly challenging. One proposed answer to this challenge is to aim at data consistency through transformations on service level [5][6].

B. *Related work*

Various standard development activities have focused on the topic of schema transformation. These include the work of the Open Geospatial Consortium (OGC) on the concept of Translating Web Feature Service [7] and the results of the ORCHESTRA project as the specification for service type called schema mapping service [8]. The recent examples of the schema transformation-related standardisation include the INSPIRE specification for the generic Transformation Service, published as EC Regulation [9] and the related technical guidance on Schema Transformation Network Service [10].

The HUMBOLDT project has worked extensively on developing mechanisms for schema transformation. The results of this work include a specification for a Web Service called Conceptual Schema Transformer and the development of an interactive schema mapping tool called HALE (HUMBOLDT Alignment Editor) [11].

The problem of schema transformation in the European-wide spatial data provision framework has been studied by Friis-Christensen et al. [12]. Donaubauer et al. describes a mechanism for supporting schema transformations as an extension to the OGC Web Feature Service (WFS), called model-driven WFS (mdWFS) [13]. Curtis and Müller describe a categorisation of atomic schema transformation components in the context of relational database to Extensible Markup Language (XML) translations [14].

 ISBN: 978-1-61208-178-6

Recent work related to schema transformation issues include the paper of Wiemann et al. on schema transformation as a component in the SDI development [15] and the work of Foerster et al. on combining schema transformation with another process, model generalisation, in a Web processing work flow [16].

C. *Structure of the paper*

The work described in this paper has been carried out in the context of a major EU-funded development project ESDIN (European Spatial Data Infrastructure Network), aimed at helping the National Mapping and Cadastral Agencies (NMCAs) to attain compliance with INSPIRE regulations [17].

In Chapter II the concept of schema transformation is elaborated and some further background information given. In Chapter III the alternative architectures for schema transformation in the Web Service environment are described. The Chapter IV focuses on the results achieved in the implementation of the approach in the context of a major EU development project. The paper ends with a conclusion in Chapter V.

II. SCHEMA TRANSFORMATION CONCEPTS

A. *Transformation Categories*

Content transformations dealing with geospatial data can be considered in various levels of abstraction [18]. Transformations that only consider pure *syntactic harmonisation* of content have a well-established tradition and are routinely performed in the existing and upcoming SDIs. Typical examples of these transformations include the use of Geography Markup Language (GML) as the data encoding standard and the internal processes of the Web Feature Service (WFS) implementations that transform data content from the internal representation of the used data store to the GML-encoded output. Integration of two heterogeneous data stores becomes thus technically possible, but as semantic differences are not considered, the combined data set becomes practically useless.

More advanced transformations take into account the *semantics* of the source data and aim at matching semantically equal or close-to-equal data items to each other. The resulting combined data set is internally consistent and can be accessed using a single semantic framework. In a more advanced approach the semantic correspondence between the data items can be indicated in a more flexible way, facilitating the use of semantic similarity as a measure for controlling the data matching process. Ontologies are sometimes used as a tool for managing flexible semantics-aware transformations [19].

The transformation might focus predominantly on restructuring the data content and applying new naming (typically in a different language) to the constructs that encompass individual data items. This kind of data transformation is called *schema transformation*, as the most typical scenario for this process is transforming spatial data set from one schema (for instance a national model) to another (for instance a common European level model). The transformation is expected to maintain the semantics of the data content as far as possible.

Due to the special nature of geospatial data, various *geometric transformations* are essential. An advanced schema transformation may involve geometric operations as a component. These might involve change of the type of the geometric primitive used to depict the position of a spatial object, for instance transforming a small polygon to a point or an elongated polygon to a linear primitive.

B. *Schema Mapping*

An essential part of the transformation process is the establishment of the mapping between data items in the source schema and the corresponding items in the target schema. This mapping can be defined either on the abstract, conceptual level or on a concrete, implementation-dependent schema level. Some research papers recommend the use of schema mapping on conceptual level [20]. The benefits that this approach yields can be listed as follows:

- The person defining the mapping does not need to be knowledgeable on the implementation level details
- The same mapping, done on the conceptual level, can be subsequently implemented in various technical frameworks
- The mapping made on conceptual level can be expected to remain valid for longer period of time when compared with a mapping based on rather fast-changing technology solutions

The weaknesses of this approach include:

- A separate processing step is required to make the conceptual mapping usable in a concrete implementation environment, often requiring manual intervention
- There is no widely adopted mechanism available for recording the conceptual level schema mapping

The technologies proposed as the method for encoding schema mapping on conceptual level include the OMG (Object Management Group) standard for model transformation MOF 2.0 Query/View/Transformation [21] and Ontology Mapping Language (OML). For instance the Geo Ontology Mapping Language (GOML) specified in the HUMBOLDT project is based on OML [22].

III. ARCHITECTURAL CONSIDERATIONS

There are three different approaches for organizing the schema transformation in relation to the Web service-based data delivery workflow.

Firstly, the transformation can be carried out as a completely offline pre-processing step, typically performed while transferring a data set from the original data store to a service database. This approach is most appropriate in cases when the difference between the native schema and the output schema is so significant that the transformation requires long processing time or even manual interventions. In some cases the original data store might be organized in suboptimal way from the efficient service-based content delivery point of view, thus making the use of a separate service database a necessity. This is often also justified from the security con-

 ISBN: 978-1-61208-178-6

cerns' point of view. However, in some cases it might be possible to maintain the transformed data content inside the same database management system with the source data set, for instance by applying mechanisms like database views and triggers.

Secondly, the transformation can be organized as an online process carried out during the interactive request-response dialogue in the Web Service environment. In this approach the transformation is part of the data delivery workflow and has to be performed as an on-the-fly process. The transformation is carried out only on a subset of the data content, retrieved from the database on the basis of the query sentence of the data request. Being a real-time process the on-the-fly transformation must be straightforward enough to be performed as a fully automatic step in a synchronous transaction.

In some cases the right solution would be to combine the two above-mentioned approaches. A separate transformation process is carried out while transferring data from the production environment to the delivery database, accessible by the online services. This transformation is performed as a pre-processing step outside of the online service environment. In addition to this major batch-oriented transformation, the system might still make use of the on-the-fly transformation approach by also including transformation functionality into the online delivery process. This latter transformation will take care of all the remaining modifications that are still needed to make the delivered data set fully compliant with the target schema. The various approaches for including a transformation into the data delivery workflow are depicted in the Fig. 1.

In an operational service-oriented spatial data delivery environment schema transformation can be organised in various different ways. In the following the main approaches are explained, together with a brief analysis of their strengths and weaknesses. The framework has been initially defined in the context of the service development work of the ESDIN project [23].

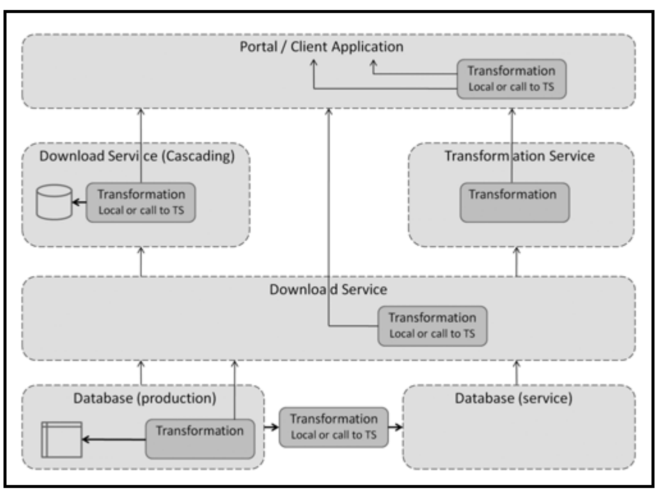

Figure 1. Transformation alternatives. The bold arrows depict offline processes. TS: Transformation Service.

A. *Database-Centric*

On database level simple on-the-fly transformations could be configured as database views. However, in most cases they will not be able to provide the functionality required for achieving full compliancy with the target schema.

In the offline approach various data schemas are maintained inside a single database. Mechanisms available for the transformation include database views, materialized views and tables that are updated by triggers on the source data tables. The transformations might be configured as a set of database scripts (e.g., PL/PGSQL scripts) and can be performed as an integral part of updates to the source data (change propagation).

B. *Service Database-Oriented*

In this approach the data provider decides to set up a separate service database containing data in the target schema. This arrangement can be motivated by the fact that the existing data maintenance platform does not support effective enough data retrieval mechanisms, or does not allow connection with the selected content access service implementation. Security-related concerns might also be the deciding factor.

The transformation from the source schema to the delivery schema is carried out as part of the batch process that transmits data content from the source database to the service database. This process can be based for instance on SQL scripts. Even several separate service databases could be set up, each of them potentially providing content for different services and supporting different schemas. The batch process carrying out the data transmission can be run periodically or it might be triggered for instance by the updates on the source data tables.

C. *Download Service Internal*

One approach for transformations in the currently existing software solutions is to apply transformations as an internal function of the content access service itself. As an example some of the currently available Web Feature Service (WFS) implementations support Coordinate Reference System (CRS) transformations as an integral part of the data access transaction. In a similar way, some WFS products support configurable schema transformation functions to be introduced to the data delivery process, so that the data items of the internal database schema can be mapped to a desired output schema. The WFS implementations that support more extensive transformations are called Translating WFS servers or Transforming WFS servers [24].

D. *Middleware Approaches*

If the transformation is performed somewhere between the source data service and the client application, the approach can be taken as an example of a middleware solution. In this case the transformation gets the input from the data service and provides the transformed data set as an output to be consumed by the user application. Two concrete approaches can be identified for this alternative: a cascading-transforming data service and a dedicated transformation service.

 ISBN: 978-1-61208-178-6

1) Cascading-Transforming Data Service

The concept of cascading services is recognised as a solution for content integration in the geospatial services development community. In this approach a service node works as a client for another service and provides the content of that service as part of its offerings to the real client applications. This way a single WFS node can provide access to resources of various individual WFS services as cascaded content, together with the resources it might be serving from its local data stores. As part of the cascading functionality the WFS can also perform schema transformations.

2) Transformation Service

The transformation could also be performed by a dedicated transformation service. In this approach the transformation service is connected to the source data service for input data and provides the transformed data as an output data set to be consumed by the user application. Unlike in the case of the cascading-transforming data service, the transformation service does not expose itself to the calling applications as content access service, but as a geospatial processing service, enabling the transformation process to be configured by calling applications.

E. Portal-Centric

The predominant role of the portal in an SDI is to integrate distributed services into a single access point. As such a portal is good candidate for also performing other processing tasks, like data transformations. A portal might contain such transforming process as an internal function or rely for the transformation on an outside resource like a transformation service.

IV. CASE IMPLEMENTATION

The ESDIN (European Spatial Data Infrastructure Network) project was a Best Practice Network project in the EU eContentplus framework program [17]. The main goal of the ESDIN project was to help the European National Mapping and Cadastral Agencies (NMCAs) in their efforts to fulfil the requirements set by the INSPIRE process. The project Consortium membership included NMCAs from 11 different European countries and it was coordinated by EuroGeographics [25]. The ESDIN project was finished in March 2011.

As one of its main goals the project focused on a coordinated establishment of INSPIRE-compliant Download Services. The services are built according to the service category 'Direct Access Download Service', thus implementing the OGC Web Feature Service (WFS) interface. The output data are provided in INSPIRE schemas, encoded in GML version 3.2.1 according to the requirements set in INSPIRE.

One of the primary areas of investigation in the project was to find out, how schema transformations from the national data models to the INSPIRE schemas could be carried out. This included tests both on off-line and on-the-fly transformation approaches. In the final stage, roughly half of the developed services follow at least partially the on-the-fly schema transformation approach, whereas the other half rely on offline processes [26].

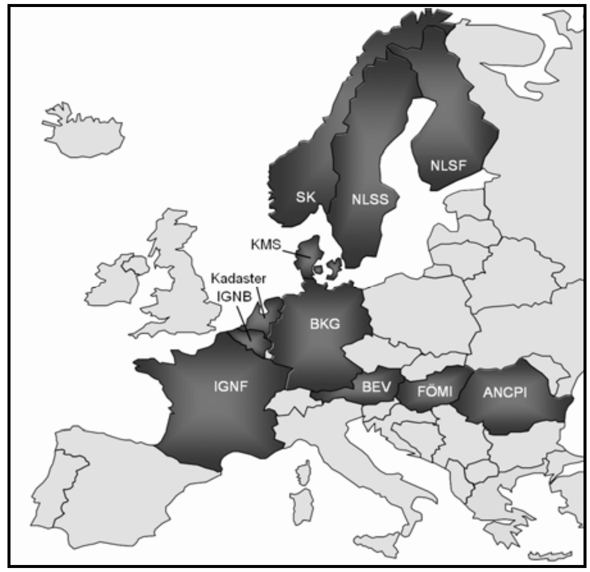

Figure 2. ESDIN countries providing content for the Download Services. Acronyms refer to the NMCA of the country.

As the main result of the service development efforts of the project, roughly 50 data sets are now available as INSPIRE-compliant Download Services. The offering cover five different data themes from the INSPIRE Annex I and has either complete national coverage or is restricted to selected cross-border test areas. The countries providing content for the ESDIN services are depicted in Fig. 2.

In the ESDIN context the proposed approach for schema mapping is to apply an SQL-like pseudo code for the purpose. The proposal includes a method for recording the mapping in a spreadsheet [27].

In the ESDIN project the NMCAs tested various different approaches for schema transformation. The architectural alternatives tested in the project are depicted with numbers 1-4 in the Fig. 3. The four cases are briefly described in the following.

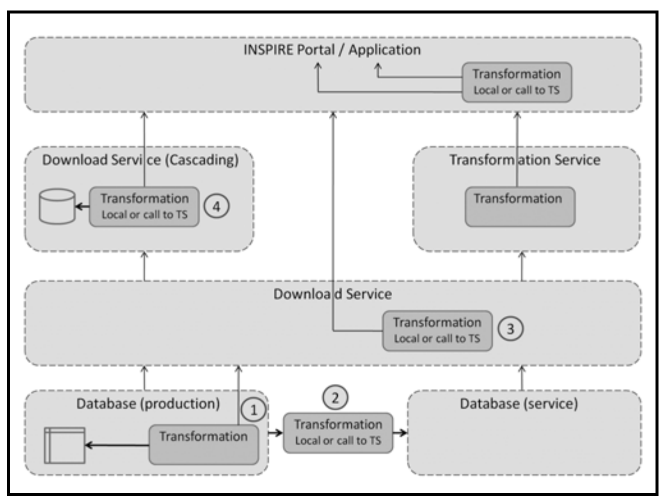

Figure 3. The four transformation architecture alternatives used in the ESDIN project. Numbers refer to cases explained below.

 ISBN: 978-1-61208-178-6

Case 1. On-the-fly transformation by database views. In this approach database views are created to transform some aspects of the data to conform to INSPIRE requirements. This approach was implemented by one NMCA.

Case 2. Offline transformation. The process steps include: extract data from the production database, transform it to INSPIRE compliant form, upload it to the service database. This approach follows the general Extract/Transform/Load (ETL) processing model. The transformation tools used include database scripting, XSLT scripting and commercial tools like FME [28][29]. This method was tested by eight NMCAs.

Case 3. Internal on-the-fly transformation carried out by Download Service. In this approach the used Download Service implementation provides support for schema transformations. These can be configured for instance by annotated XML Schema files, XSLT declarations, or by implementation-specific mechanisms. The tools used include deegree and XtraServer [30][31]. This transformation approach was implemented by eight NMCAs.

Case 4. On-the-fly transformation carried out by a cascading-transforming Download Service node. In this alternative the original Download Service provides data in local schema. This service is accessed by another Download Service that uses the first service as its data source. During the data request the cascading-transforming Download Service carries out schema transformations, typically both on query and data streams. This approach was tested by two NMCAs.

V. CONCLUSION

The experiences gained in the ESDIN service development suggest that many of the transformation types required for INSPIRE compliancy are too complicated or time consuming for the on-the-fly processing approach. Roughly half (12) of the developed services (22) are (at least partially) based on on-the-fly transformation principles. Only eight services are based on pure on-the-fly transformation approach. As a result it can be concluded that on-the-fly transformations did not take as central a role in the project as initially expected. On the other hand, the combined approach, in which both offline and on-the-fly transformations are used together, seems to be rather promising [26]. The final service configuration of the ESDIN project is illustrated in Fig. 4.

A database level solution was raised during the project experiments as a new kind of alternative for content transformations. The concrete implementations developed in the project are based on the use of database views and thus represent on-the-fly approach on schema transformation on the database level.

The experience gained in the ESDIN project suggests that various different approaches for content transformation can be adopted with successful results. It seems that general recommendation cannot be given on any single transformation method, as the best approach always depends on local conditions. One of the most important deciding factors is, how far the national schema is from the related INSPIRE schema. The bigger the difference, the less probable it is that on-the-fly solutions would yield acceptable results.

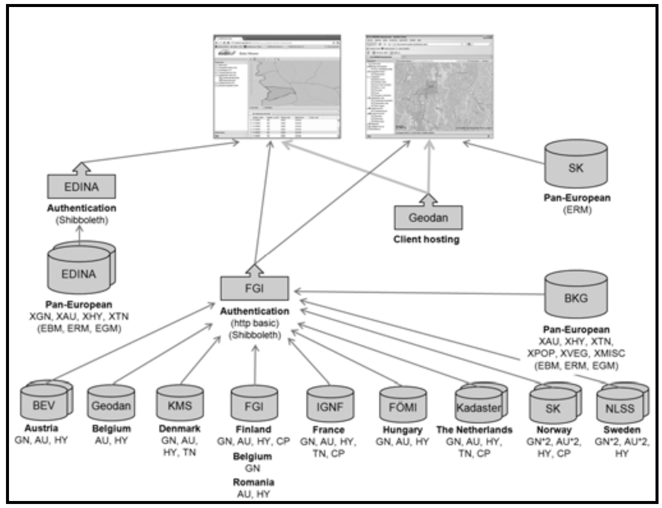

Figure 4. Final service configuration of the ESDIN project. Acronyms below the country names indicate the INSPIRE themes supported. Themes starting with 'X' denote schemas developed in the project as extensions to the INSPIRE schemas.

A standardised way to describe the schema transformation is a useful tool for communication among the individuals involved in the schema mapping process. The mechanisms proposed by the ESDIN project have been found to be helpful in this respect.

The number of transformation steps has to be kept to minimum, because each transformation process must be first developed, and further on maintained too. Every transformation also potentially introduces a certain amount of error or information loss to the process.

In most of the cases the off-line approach for schema transformation seems to provide best results. On-the-fly methods can be applied for fine-tuning minor schema nuances or in cases where the differences between the local schema and the corresponding INSPIRE schema are minor. The most prominent advantage in using the on-the-fly approach is the direct connection to up-to-date content.

The main disadvantage in the offline transformation method is the need to have a separate service database and the inherent problem of keeping it up-to-date. On the other hand, this approach also supports easy fusion of the data with other source data sets. Security concerns can also be better taken care of.

The task of setting up an INSPIRE-compliant Download Service is not straightforward. Reasonable amount of resources have to be invested in the process. The main issue to resolve is the transformation of content stored in the national schema to be compliant with INSPIRE schemas. This was successfully tackled in the ESDIN project with solutions developed for both offline and on-the-fly approaches.

ACKNOWLEDGEMENT

The ESDIN project was funded by European Commission in the Framework Programme eContentplus (grant agreement number ECP-2007-GEO-317008).

REFERENCES

[1] A. Friis-Christensen et al., "Building service oriented application on top of a spatial data infrastructure – A forest fire assessment example," Proc. AGILE International Conference on Geographic Information Science, 2006, Visegrad, pp. 119–127, Hungary.

[2] L. Lehto, "Real-time content transformations in the European spatial data infrastructure," Proc. International Cartographic Conference, Nov 15-21, 2009, Santiago, Chile, CD-ROM.

[3] JRC, Web site of the INSPIRE process, http://inspire.jrc.ec.europa.eu <retrieved: 11, 2011>

[4] EC, DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), at: http://eur-lex. europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML <retrieved: 11, 2011>

[5] L. Lehto, "Schema transformation as a Web service, Proc. Nordic GIS Conference, Helsinki, Oct 2-4, 2006, pp. 39-44.

[6] L. Lehto, "Schema translations in a web service-based SDI," Proc. AGILE International Conference on Geographic Information Science, 'The European Information Society: Leading the way with geo-information', May 8-11, 2007, Aalborg, Denmark..

[7] Y. Bishr and K. Buehler, “GOS Transportation Portal, Implementation Architecture and Lessons Learned,” Internal OGC Interoperability Program Report.

[8] M. Lutz, "Specification of the schema mapping service," OA services specifications, Orchestra project, Deliverable D3.4.3, Joint Research Centre of the EU Commission.

[9] EC, “Commission Regulation (EU) No 1088/2010 of 23 November 2010, amending Commission Regulation (EC) No 976/2009 as regards download services and transformation services,” at: http://eur-lex.europa.eu/LexUriServ/ LexUriServ.do?uri=CONSLEG:2009R0976:20101228:EN:PDF <retrieved: 11, 2011>

[10] INSPIRE, “Technical Guidance for the INSPIRE Schema Transformation Network Service,” at: http://inspire.jrc.ec.europa.eu/documents/Network_Services/JRC_INSPIRE-TransformService_TG_v3-0.pdf <retrieved: 11, 2011>

[11] T. Reitz, M. de Vries, and D Fitzner, “Conceptual Schema Specification and Mapping”, Deliverable A7.0-D3, HUMBOLDT project, at: http://www.esdi-humboldt.eu/press/public_deliverables.html <retrieved: 11, 2011>

[12] A. Friis-Christensen, S. Schade, and S. Peedell, “Approaches to solve schema heterogeneity at the European level,” Proc. EC-GI & GIS Workshop, Alghero, Sardinia, Jun 29 – Jul 1, 2005.

[13] A. Donaubauer, F. Straub, and M. Schilcher, "mdWFS: a concept of webenabling semantic transformation," Proc. AGILE International Conference on Geographic Information Science, May 8-11, 2007, Aalborg, Denmark.

[14] E. Curtis and H. Müller, “Schema translation in practice,” Snowflake Software White Paper, at: http://www.snowflakesoftware.co.uk/news/papers.htm <retrieved: 11, 2011>

[15] S. Wiemann et al., "Web services for spatial data transformation and exchanges in SDI: a prototypical implementation of the LPIS quality assurance test bed services," International Journal of Spatial Data Infrastructure Research, unpublished.

[16] T. Foerster, L. Lehto, T. Sarjakoski, L. T. Sarjakoski, and J. Stoter, 2010. "Map generalization and schema transformation of geospatial data combined in a web service context," Computers, Environment and Urban Systems, 34(1): 79-88, at: http://dx.doi.org/10.1016/j.compenvurbsys.2009.06.003. <retrieved: 11, 2011>

[17] ESDIN, European Spatial Data Infrastructure Network, project Web site, at: http://www.esdin.eu <retrieved: 11, 2011>

[18] L. Lehto, "Real-Time Content Transformations in a Web Service-Based Delivery Architecture for Geographic Information," Doctoral dissertation, Helsinki University of Technology. Publications of the Finnish Geodetic Institute, N:o 138, 2007, Kirkkonummi, 51 p. + 99 p.

[19] F. Fonseca, G. Câmara, and A. M. Monteiro, "A Framework for Measuring the Interoperability of Geo-Ontologies," Spatial Cognition & Computation, Vol. 6, No. 4, 2006.

[20] H. R. Gnägi, A. Morf, and P. Staub, "Semantic Interoperability through the Definition of Conceptual Model Transformations," Proc. AGILE Conference on Geographic Information Science, 2006, Visegrád.

[21] OMG, Meta Object Facility (MOF) Query/View/Transformation specification (QVT) http://www.omg.org/ spec/QVT/ <retrieved: 11, 2011>

[22] HUMBOLDT, Web site of the Humboldt project, at http://www.esdi-humboldt.eu/home.html <retrieved: 11, 2011>

[23] L. Lehto (ed.), "Best practice for content transformations enabling INSPIRE-compliant data delivery," Deliverable 11.1, Public report, ESDIN project, European Spatial Data Infrastructure Network, Oct 9, 2009, ECP-2007-GEO-317008, 44 p.

[24] E. Curtis, "Schema Translation and Semantics in Data Interoperability," Snowflake Software White Paper, at: http://www.snowflakesoftware.co.uk/news/papers.htm <retrieved: 11, 2011>

[25] EuroGeographics, Association of National Mapping, Land Registry and Cadastral Agencies, Web site at: http://www.eurogeographics.org <retrieved: 11, 2011>

[26] L. Lehto (ed.), "Recommendations for operational deployment of services," Deliverable 11.5, Public report, ESDIN, European Spatial Data Infrastructure Network, Feb 28, 2011, ECP-2007-GEO-317008, 67 p.

[27] L. Lehto and F. Nissen, "Schema transformations in the European spatial data infrastructure," Proc. INSPIRE Conference, 22-25 June, 2010, Krakow, Poland. Poster.

[28] W3C, Extensible Stylesheet Language Transformations (XSLT), at: http://www.w3.org/TR/1999/REC-xslt-19991116 <retrieved: 11, 2011>

[29] Safe Software, The FME Techonology, at: http://www.safe.com/fme/fme-technology/ <retrieved: 11, 2011>

[30] deegree, the web site of the deegree community, at: http://www.deegree.org <retrieved: 11, 2011>

[31] interactive instruments, XtraServer product information (in German language), at: http://www.interactive-instruments.de/index.php?id=xtraserver&L=1 <retrieved: 11, 2011>

 ISBN: 978-1-61208-178-6