A Workflow for Analyzing Interoperability of Geospatial Data Sets at Different Administrative Levels – A Case Study in Forest Fire Data

Kaori Otsu[a], Sven Schade[b], Laura Díaz[c]

[a]Knowledge Engineering Group, University Jaume I, Castellón, Spain
kaori.otsu@uji.es
[b]Institute for Environment and Sustainability, European Commission, Joint Research Centre
Ispra, Italy
sven.schade@jrc.ec.europa.eu
[c]Institute of New Imaging Technologies, University Jaume I, Castellón, Spain
laura.diaz@uji.es

ABSTRACT

The increasing demand for integrated assessment in the environmental sciences makes it crucial to understand geospatial data sets and to increase their sharing at different administrative scales. To address these issues we propose a workflow to analyze data interoperability from a regional level to a pan-European level. Within a case study in forestry, upward interoperability of burned area data from the Valencia Community in Spain and Europe are examined on the syntactic, schematic and semantic dimension. Apart from revealing practical harmonization issues of today, our work resulted in a workflow for interoperability analysis of geospatial data sets. This may be used to judge the feasibility of data interoperability in any project, which aims to use geospatial data from multiple sources.

1 INTRODUCTION

Over the past two decades the distribution of geospatial data has significantly increased as information technologies advanced (Masser 2005). Since data sets often derive from different sources, it is necessary to establish a common framework for data sharing and exchange. The common framework can be designed in a spatial data infrastructure (SDI) where geospatial data can be readily accessible in cooperation with various stakeholders through agreed policies and common standards (Phillips et al. 1998).

Taking an example from forest fire data, at the European level, the European Forest Fire Information System (EFFIS)[1] is implemented in compliance with the guidelines of the Infrastructure for Spatial Information in Europe (INSPIRE 2007). At national level, Spain is nowadays adopting INSPIRE at different administrative levels. The national SDI (IDEE)[2] serves the central node connected to the directive. While IDEE contains basic forest cover data, the Ministry of Environment and Rural and Marine (MMA)[3] intends to allow forest fire data managed by the national forestry program accessible in the IDEE (MMA 2009). At regional level, forest fire data collected from autonomous regions are accessible through the forestry program (MMA 2009).

Motivated by the EU-funded EuroGEOSS project (EuroGEOSS 2009), we investigated the interoperability of forestry data sets at different administrative levels. We found it useful to compare and analyze a common area of interest (i.e. a specific autonomous region in Spain) on regional, national, and European scales. This helps to overcome heterogeneous geospatial data at different levels of detail to be commonly used in forest fire models for prediction and propagation.

The remainder of this paper is structured as follows. Key concepts related to interoperability of geospatial data are provided in the next section. With a use case in forest fire data, we introduce the workflow for interoperability analysis (section 3). We generalize over our findings in section 4.

[1] http://effis.jrc.ec.europa.eu/
[2] http://www.idee.es
[3] http://www.mma.es/portal/secciones/biodiversidad/banco_datos

2 BACKGROUND

This section contains a brief overview of central notions related to data interoperability. Relevant formats, standards and tools are presented. Furthermore, we introduce ontologies as a means to address semantic aspects in data model (aka schema) matching and data transformation.

2.1 Data Interoperability in the SDI Context

In the context of SDI, interoperability is the ability to exchange and manipulate geospatial resources across distributed systems without having to consider the heterogeneous format of the source (Bishr 1998). Several kinds of interoperability should be considered when discussing interplay of information systems. From the kinds defined by the European Interoperability Framework (EIF) (European Commission 2010) we restrict our research to data interoperability and aspects of data models.

We first focus on interoperability on the syntactical dimension by assessing the availability of standardized components at server or client side. *Syntactic interoperability* requires the integration of elements from various systems such as data formats and standards. Within the geospatial domain, interoperability is ensured by efforts most prominently by ISO/TC 211 and Open Geospatial Consortium (OGC). The OGC has proposed a number of standards promoting syntactic interoperability of geospatial data sets, as well as harmonized data access as we will see in section 3 when analyzing interoperability.

Schematic and semantic interoperability is analyzed by adopting techniques for data transformation. *Schematic interoperability* is described by common classifications and hierarchical structures while *semantic interoperability* harmonizes meanings of terms. They can be improved by using metadata standards, schemas and ontologies (Bishr 1998).

2.2 Ontologies and Geospatial Data Transformation

Common software solutions do not provide means for identifying schematic and semantic incompatibilities. We suggest applying ontology-based schema matching and the definition of transformation rules for assessing schematic and semantic interoperability. For us, an ontology is 'an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words' (Guarino 1998).

Following previous work (Friis-Christensen et al. 2005), we use semantic descriptions of schema elements to infer relations between attributes of the source and target data models. We use ontologies in order to formalize the attributes of data models and domain dependent categorizations, such as forest classifications, and inferring relations between the categories that are used in the source and target schemas.

Once matches are identified, various Extract, Transform and Load (ETL) tools can be used for executing them on a given source data set. These include the Feature Manipulation Engine (FME), GoPublisher, Spatial Data Integrator (SDI), and GeoXSLT (Chunyuan et al. 2010).

3 WORKFLOW FOR INTEROPERABILITY ANALYSIS WITH A USE CASE IN FOREST FIRE DATA

In this section we introduce a workflow for analyzing interoperability of geospatial data sets and apply it to forest fire data from the EFFIS and the Ministry of Environment, Water, City Planning and Housing of Valencia Community[4] in Spain. National forest fire data in Spain are collected from autonomous regions; therefore, regional data from CMA are directly tested for interoperability with the EFFIS data. Figure 1 illustrates the two parts of the workflow for interoperability analysis of forest fire data between EFFIS and CMA. In the syntactic approach (left part of the figure), we attempt to perform Geographic Information System (GIS) overlay analysis of the two data sets. With the combination of schematic and semantic approaches (right part of the figure), schema matching is tested from the CMA data model to the EFFIS data model by the use of ontologies. Depending on the testing purpose, these approaches on syntactic, schematic and semantic dimensions can be applied for interoperability analysis independently. The reminder of section demonstrates our tasks for interoperability analysis in steps according to Figure 1.

[4] http://www.cma.gva.es/intro.htm

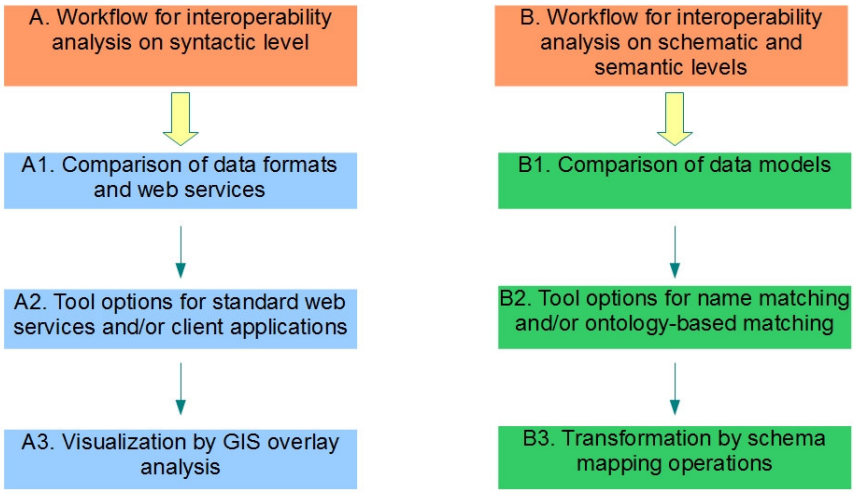

Figure 1: Workflow for interoperability analysis of geospatial data sets.

3.1 Tasks for Analyzing Syntactic Interoperability of Forest Data Sets

Differences in data formats and the accessibility via web services are tested for syntactic interoperability by performing overlay analysis. We suggest visualizing an overlay image of the two geospatial data sets using a common GIS. The integration into the GIS software requires data set encoding in a common format and accessibility of the data set.

Since we work in a distributed environment, we distinguish server-side tools from client-side tools. If data sets are shared via web services with standard interfaces and using common data encodings, they can be displayed in a single client view. Web services interfaces such as OGC Web Mapping Service (WMS), Web Feature Service (WFS), and Web Coverage Service (WCS) are common in SDIs, providing web links for viewing and downloading geospatial data sets in a standardized way. Such data sets can be consumed via web portals or GIS applications such as ArcGIS[5] and gvSIG[6], regardless of the original format of the source. These applications can make geospatial data available in original formats or available from web services to be visualized on the fly. Thus, we have to consider three options how two data sets can be displayed in the same client view:

1) Both data sets are available via standard web services, including common encoding.
2) Both data sets are provided locally as created or downloaded files, in a format that can be added onto a GIS client application.
3) One data set is stored locally (in the original format) while the other is available from standard web services.

3.1.1 Comparison of Geospatial Data Formats and Access

As the first step in the syntactic approach (A1 in Figure 1), the two data sets are compared for data formats and web services in Table 1. EFFIS provides a web entry point (map viewer) to display the image of burned area data directly on the addressed web page. To view the same data outside the EFFIS web page, data from the current year are accessible via WMS. However, data from previous years are available upon request (in SHAPE[7] format). The second source, burned area data from CMA, is accessible via ArcIMS, which is connected to an internal network[8], but not to the (public) World Wide Web.

[5] http://www.esri.com/software/arcgis/index.html
[6] http://www.gvsiggva.es
[7] The most common format for geospatial vector data introduced by ESRI.
[8] http://intranet.cma.gva.es

Table 1: Comparison of forest fire data between regional and European levels.

SDI Level	Regional	European
Area	Valencia Community	Europe
Layer	Forest Fires (incendios)	Burned Area
Access	ArcIMS (internal)	Map Viewer
Format	PNG (SHAPE upon request)	PNG (SHAPE upon request)
Metadata	ISO 19115	ISO 19115
Spatial resolution	20 m	250 m
Minimum fire size	0.05 ha	50 ha
Temporal resolution	annually	annually
Spatial reference system	ED50 / UTM Zone 30N	ETRS89 / ETRS-LAEA
Projection	EPSG:23030	EPSG:3035
Language	Spanish or Valencia	English
Owner	CMA	EFFIS

3.1.2 Assessing Options for Viewing the Geospatial Data Sets

In the next step (A2 in Figure 1), options to view these two data sets in the same client are analyzed. Since ArcGIS view and download services are not OGC standard web services, accessibility to burned area data provided by CMA is limited. Forestry domain experts who work at CMA could add CMA data via ArcIMS and EFFIS data in SHAPE format, using ArcGIS (corresponding to option 3). As we do not have access to the CMA internal network, we need to request the data from CMA, for example in SHAPE format, and to display it in ArcGIS together with the EFFIS data set, also in SHAPE format (corresponding to option 2).

3.1.3 Overlaying the Geospatial Data Sets

Lastly (A3 in Figure 1), visualization by GIS overlay is realized as shown in Figure 2, region Depending on the level of detail contained in each data set, overlay analysis can also address scale issues. It is apparent that CMA data shows more detailed matching along the boundary of region 'A'. Another scale issue illustrates that some small burned areas (regions 'B', 'C', and 'D') mapped by CMA are missing in the EFFIS data set.

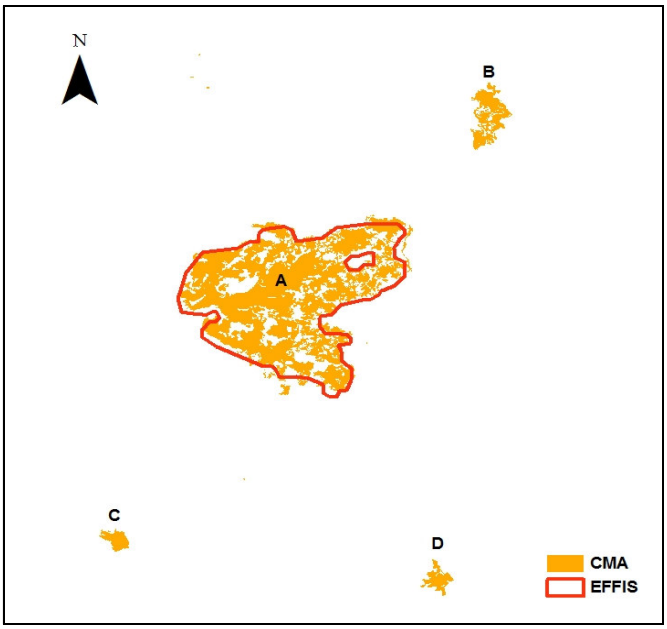

Figure 2: GIS overlay image of burned areas in Valencia Community mapped by CMA (solid region) and EFFIS (solid line) at scale 1:200 000.

3.2 Tasks for Analyzing Schematic and Semantic Interoperability

Schematic and semantic interoperability can be tested by comparing involved schemas and the descriptions used for the schema elements, such as object types and attributes. ETL tools provide functionality for directed schema matching, i.e. from a source to a target. However, this approach is not always sufficient to find schema matches and may even introduce conceptual mismatches (Reitz 2010). Ontologies aid to conceptually analyse the schemas based on potential matches and also provide a possibility to determine shared conceptualizations between source and target attributes where they were not potentially matched on the schematic dimension (Schade 2010). Accordingly, we consider that two data models can be matched by the following options:

1) Source and target attributes are mapped by name matching.
2) Source and target attributes are mapped by ontology-based matching.
3) Source and target attributes are mapped by a combination of name matching and ontology-based matching.

3.2.1 Comparison of Geospatial Data Models

Firstly (B1 in Figure 1), the two data models are compared. The source data model represents burned areas caused by forest fires in Valencia Community. Burned areas are categorized into non-wooded and wooded forest surfaces. The target data model by EFFIS represents burned areas damaged by fires in Europe, including non-forest cover types such as agricultural and artificial areas. Considering these differences in the two data models, our intent is to test the third option by combining the schematic approach (option 1) and the semantic approach (option 2) to achieve more or better matching candidates.

3.2.2 Matching Attributes of the Geospatial Data Models

The next step (B2 in Figure 1) starts with name matching where attribute names in the source data model need to be translated from Spanish to English. Different terms used in different languages can be translated by referring to a multi-language dictionaries or thesauri and exploiting short-forms, acronyms, synonyms, and hypernyms (e.g. Tree and Pine) (Rahm and Bernstein 2001). With this approach, five attributes were manually matched to the names of target attributes in Figure 3.

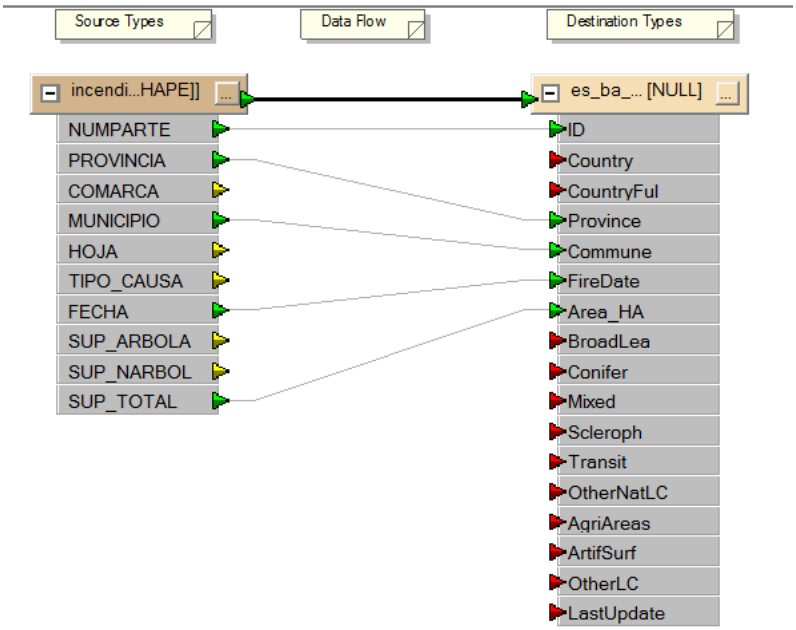

Figure 3: Name matching attributes from source to target data model using the FME.

In order to perform semantic matching, we follow previous work (Friis-Christensen et al. 2005) and apply a (semantic) classification matching approach with an ontology editor Protégé. We propose to establish two application ontologies based on CMA and EFFIS, followed by semantic analysis of

the data models according to discussions with forestry domain experts. Data standards for Forest Inventory in Valencia Community by CMA and CORINE Land Cover Classification by EFFIS were used as guidelines. Using the reasoner in Protégé, equivalent and similar classes were reclassified, which enabled the five name matching attributes to be further tested on the semantic dimension. Additionally, the following ontology-based matching attributes were identified and inferred as similar classes in Figure 4:

- SUP_NARBOL ≈ subclassOf (Transit)
- SUP_ARBOLA ≈ superclassOf (Forests: Conifer/BroadLea/Mixed)

We revised the transformation rules to add these new matching attribute in the Figure 3, mapping 'SUP_ARBOLA' to 'Conifer' and 'SUP_NARBOL' to 'Transit', respectively.

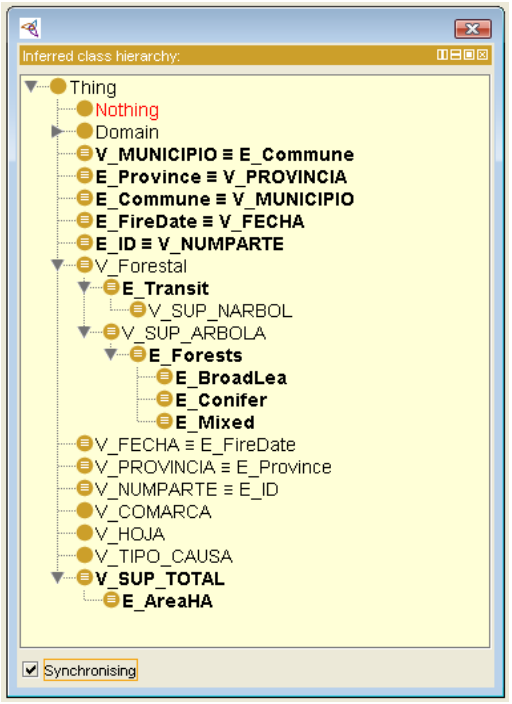

Figure 4: Equivalent (≡) and similar (sub/super) classes inferred by reasoner in Protégé for name matching attributes.

3.2.3 Transforming Geospatial Data Sets

Once information about attribute matches has been gathered, it can be used to define the actual transformation operations (B3 in Figure 1) that are required for harmonizing a source with a target (Chunyuan et al. 2010). As the workflow of transformation operations is illustrated in Figure 5, target attributes with green arrows were successfully transformed from source attributes. Thus we were partially able to deliver ontology-based schema mapping in transformation from the CMA to the EFFIS data model. Then mapping rules to perform this transformation were saved in the FME transformation file.

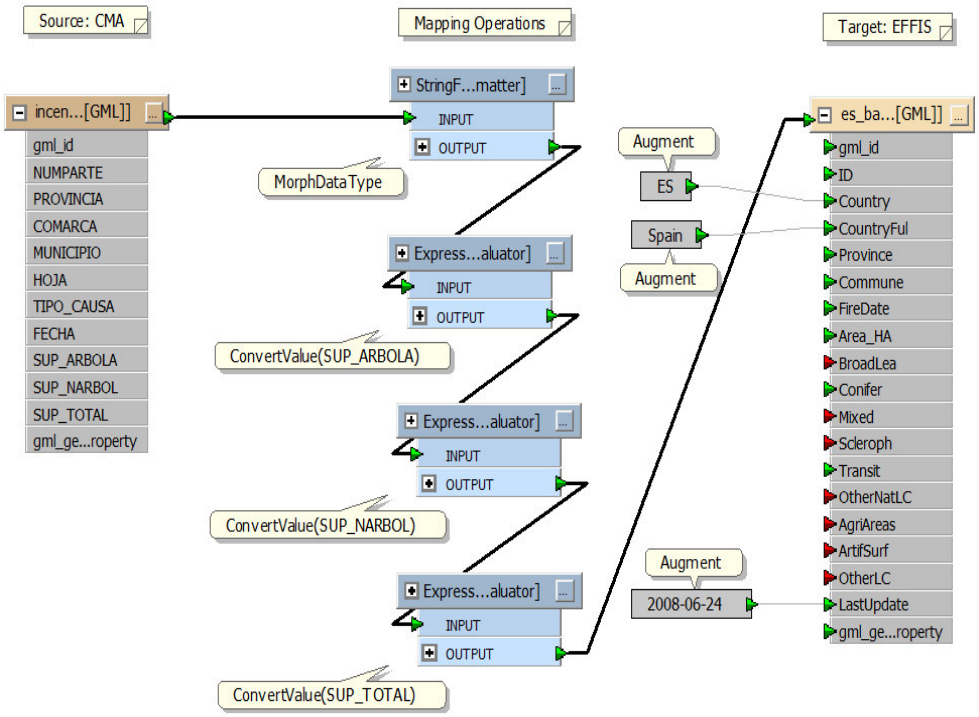

Figure 5: Mapping operations added in the FME to perform transformations from source attributes to target attributes.

4 CONCLUSIONS AND FUTURE WORK

Our case study in forestry demonstrated how heterogeneous geospatial data sets are from regional to European levels. Forest fire data from Europe and the EFFIS Member States (Valencia Community in Spain) are not syntactically compatible by server-side tools, therefore, EFFIS is expected to accelerate the process of implementing standard web services. At regional level, not only CMA should allow the link of ArcIMS publicly available, but also it can be standardized to OGC web services. Using client-side tools which allow geospatial data to be added via various options of web services can also increase the chance of achieving syntactic interoperability.

For achieving schematic interoperability, ETL tools such as FME enable various types of data transformation from source to target attributes. To identify matching attributes, name matching approach was quick and simple when source and target attributes matched linguistically. However, the linguistic approach cannot always provide sufficient information to identify schema matches. Ontological modeling helped identify a common concept between the source and target data models, especially in cases where matching attributes were not found at the schematic dimension. Even with the aid of application ontologies based on source and target data standards, identifying a common concept between them remained difficult for some attributes such as forest types. In cases where semantic common ground cannot be reached at application level, the domain ontology may enable semantic matching (Friis-Christensen et al. 2005).

Future work related to this case study includes the creation of the new domain ontology for redefining ‘forest’ as common ground according to the level of abstraction. This work can be established on the basis of existing domain and foundational ontologies, which have been introduced for schema transformation (Schade 2010). It may require top-down and bottom-up approaches in a hierarchy between the domain and application ontologies to find optimal common ground. Additionally, we may investigate how schema mapping rules are executed in the complete process of schema transformation from source to target data so that the regional data can be inputted into the forest fire model developed by EFFIS. It may require another rule language, such as the Rule Interchange Format (RIF) to reuse and exchange those mapping rules that can be processed by other execution tools.

BIBLIOGRAPHY

Bishr, Y. (1998) Overcoming the Semantic and Other Barriers to GIS Interoperability. International Journal of Geographical Information Science, 12(4), pp.299–314.

Chunyuan, C., Schade, S. and Gudiyangada, T. (2010) Schema Mapping in INSPIRE - Extensible Components for Translating Geospatial Data. Proceedings of 13th AGILE International Conference on Geographic Information Science, Guimarães, Protugal.

EuroGEOSS. (2009) D.3.1: report on user requirements for the EuroGEOSS Forestry Operating Capacity. http://www.eurogeoss.eu/Documents/ EuroGEOSS_D3-1.pdf, last date accessed: 10.2010.

European Commission. (2010) European Interoperability Framework (EIF) for European Public Services. http://ec.europa.eu/isa/strategy/doc/110113__iop_ communication_annex_eif.pdf, last data accessed: 01.2011.

Friis-Christensen, A., Schade, S. and Peedell, S. (2005) Approaches to solve schema heterogeneity at European Level. Proceedings of 11th EC-GI & GIS Workshop, ESDI: Setting the Framework, Alghero, Sardinia, Italy.

Guarino, N. (1998) Formal ontology and information systems. Proceedings of 1st International Conference on Formal Ontologies in Information Systems (FOIS), Trento, Italy.

INSPIRE. (2007) 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. http://eur-lex.europa.eu/ LexUriServ /LexUri Serv.do?uri=OJ:L:2007:108:0001:0014:en:PDF, last data accessed: 10.2010.

Masser, I. (2005) GIS Worlds: Creating Spatial Data Infrastructures. Redlands, California. ESRI Press.

MMA (Ministerio de Medio Ambiente y Medio Rural y Marino). (2009) Biodiversity: Legislation and Agreements. http://www.mma.es/portal/ secciones/biodiversidad/legislacion_convenios, last data accessed: 10.2010.

Phillips, A., Williamson, I. and Ezigbalike, C. (1999) Spatial Data Infrastructure Concepts. Australian Surveyor, 44 (1), pp.20-28.

Rahm, E. and Bernstein, P. (2001) A survey of Approaches to Automatic Schema Matching. Very Large Data Bases Journal, 10 (4), pp.334–350.

Reitz, T. (2010) A Mismatch Description Language for Conceptual Schema Mapping and Its Cartographic Representation. Geographic Information Science: Lecture Notes in Computer Science 6292, pp.204-218.

Schade, S. (2010) Ontology-Driven Translation of Geospatial Data. AKA, Heidelberg, Germany.