TOWARDS INTEROPERABILITY OF SPATIAL PLANNING DATA: 5-STEPS HARMONIZATION FRAMEWORK

Conference Paper · June 2013

DOI: 10.5593/SGEM2013/BB2.V1/S11.051

1 author:

Karel Janecka
University of West Bohemia

TOWARDS INTEROPERABILITY OF SPATIAL PLANNING DATA: 5-STEPS HARMONIZATION FRAMEWORK

Dr. Karel Janecka

Dr. Otakar Cerba

Dr. Karel Jedlicka

Dr. Jan Jezek

University of West Bohemia, **Czech Republic**

ABSTRACT

The article describes the 5-steps harmonization framework – a complex attitude to the harmonization of spatial planning data as the importance of this process is still rapidly increasing. The reason is building of spatial data infrastructures and sharing of data coming from various resources in an unified structure. The principles shown in the article are generally valid, in this particular case the 5-steps harmonization framework was tested upon the spatial planning data. The motivation for the research was primarily the European Directive Infrastructure for Spatial Information in the European Community (INSPIRE) that states that each European Union member state must provide selected data according to the INSPIRE technical implementing rules. The innovative feature of the presented 5-steps harmonization framework is using of ontology that is used for managing of heterogeneity of spatial data from various data themes. As different institutions, organizations or companies are involved e.g. in spatial planning and building of spatial data infrastructures more technical options for data harmonization realization have been explored. The first way how to practically realize the 5-steps harmonization framework was using the ontologies and Extract-Transform-Load (ETL) tools. Further research aimed at exploring of capabilities of traditional relational database management systems and their spatial extensions for harmonization. Third, the focus was also put on traditional vendors of geographic information systems (GIS) and using of world widespread GIS for realization of the 5-steps harmonization framework. Finally the article summarizes the pros and cons of various approaches that have been explored.

Keywords: harmonization, ontology, spatial database, geographic information system

1. INTRODUCTION

INSPIRE is based on the infrastructures for spatial information established and operated by the 27 Member States of the European Union. The Directive addresses 34 spatial data themes needed for environmental applications with key components specified through technical implementing rules [11]. INSPIRE directive requires that each member state must provide the spatial data in a coherent way and therefore the process of spatial data harmonization is quite crucial for implementation of this directive. There are many aspects to be considered. The principles cited in recital 6 of the INSPIRE

Directive are considered to be a general basis for developing the interoperability needs for spatial data. The first three of the five principles are to be considered to help define the data interoperability process:

- that spatial data are stored, made available and maintained at the most appropriate level,
- that it is possible to combine spatial data from different sources across the Community in a consistent way and share them between several users and applications,
- that it is possible for spatial data collected at one level of public authority to be shared between other public authorities [11].

The figure 1 provides an overview of the components relevant for data interoperability.

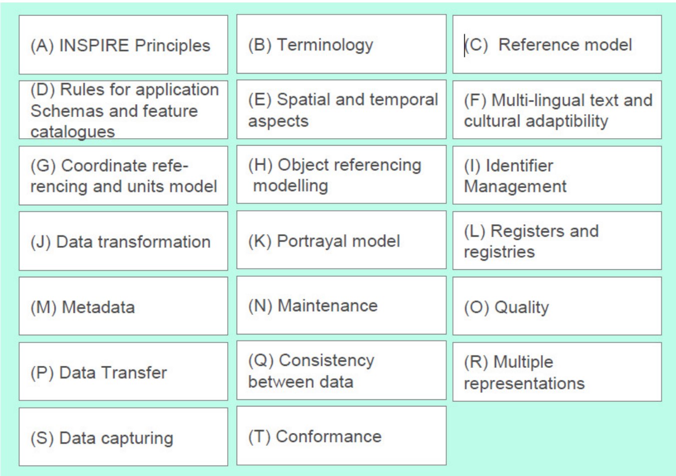

Fig. 1. Data interoperability components [8].

Each component may contribute in a different way to the interoperability of spatial data. In practice not all data interoperability components must be taken into account during the process of data harmonization. It is difficult to say if some component is more important but in every case there must be a precise understanding of both the source and target data. The right definition and description of particular harmonization steps is not possible without this exact data grasp. The key roles in this issue play domain experts, e.g. spatial planners in case of harmonization of spatial planning data. Further, the complexity of the process of spatial data harmonization also almost requires the participation of geoinformatics experts.

2. HARMONIZATION FRAMEWORK

Data harmonization is necessary for creating the possibility to combine data from heterogeneous sources (e.g. regional datasets) into integrated, consistent and unambiguous information products (e.g. European datasets). Such datasets can be then easily used with combination with other harmonized data for viewing as well as querying and analyzing. Data harmonization is complex task that has not a universal solution that will cover all possible scenarios. Ideal technical solution (system architecture, software) is always determined by many specific facts like the way how are the original data stored, size of the data and type of harmonization. Ideal solution depends also on current technical solutions used by particular organization and technical competence of IT staff [9].

The aim of the article is to present the overall harmonization framework which has been developed for harmonization of spatial planning data in scope of the INSPIRE directive. The framework is based on the knowledge and experience gained in research projects regarding the spatial planning data harmonization like Plan4all[1] (www.plan4all.eu) or Humboldt[2] (www.esdi-humboldt.eu). In practice it can be seen that the process of spatial data harmonization can be practically realized in several ways but the harmonization framework is valid for all these possible scenarios.

The proposed framework is captured on figure 2 and could be used mainly by data providers and stakeholders. The first three steps are the common steps for all scenarios. The theory of spatial data harmonization within the framework of INSPIRE and spatial planning is based mainly on [8]. The understanding both source and target data is based mainly on particular data specifications, documentation and metadata. It is quite necessary to right understand the source and target data because the process of harmonization of particular spatial planning data can be solely successful.

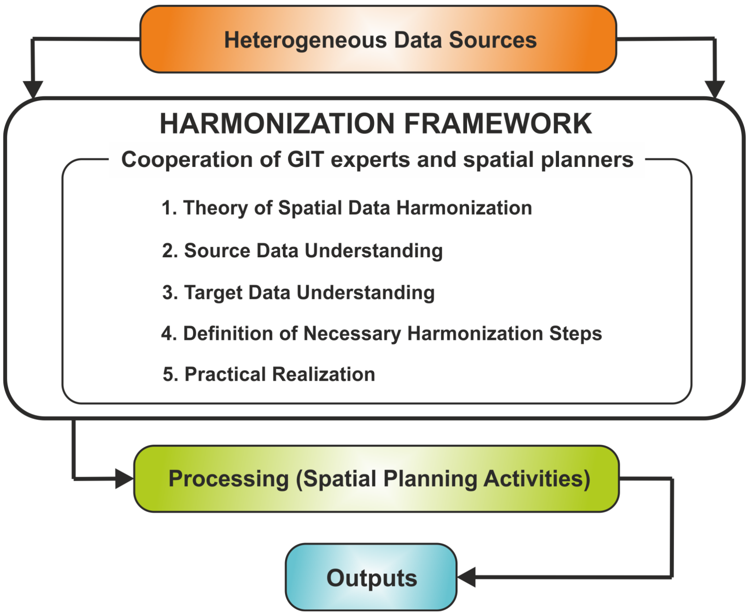

Fig. 2. The proposed 5-steps harmonization framework for harmonization of spatial planning data.

From the technical point of view there exist several ways how to handle data harmonization. One of the approaches is ETL Tools (e.g. Spatial Data Integrator, Hale). Other solution is based on usage of current capabilities provided by relation database management systems like PostgreSQL with PostGIS. Another possible solution could be using of Geographic Information System (GIS) (e.g. ArcGIS). All these possibilities were explored. In the article the main focus is put on performance and requirements of

[1] Plan4all: European Network of Best Practices for Interoperability of Spatial Planning Information.

[2] Humboldt: Towards the Harmonisation of Spatial Information in Europe.

each solution. All solutions are described in the article on particular scenarios with real datasets.

3. HARMONIZATION INSIDE A SPATIAL DATABASE

From the technical point of view the needed operations for a successful data harmonization can be done directly inside a relational database management system that has a spatial extension (e.g. PostGIS, Oracle Spatial). One can perform operations supporting the transformation of attributes or geometries that use different taxonomies or different representation of particular spatial feature. These operations can be divided into main groups:

- *filtering* – selection according to attribute or geometry predicate. For instance use just polygons with area larger than some value.
- *attribute operation* – for example renaming, change of data type, change of taxonomy (reclassification etc.).
- *geometry operation* – change the spatial representation. For example to represent polygon feature as just as point (centroid) or calculate geometry overlaps etc. Simplification or change of coordinate reference system.
- *aggregation* – calculation of new feature according to particular predicate with usage of aggregation function. For example union of all geometries according to some attribute, for instance join all features of same kind that are adjacent to each other into one new feature.

It can be found that relational database management systems can offer great capabilities to deal with this. Further it is described how to perform change of taxonomies and afterwards geometry union of features that fall under the same class.

Source data

Consider original data with simple schema as shown on figure 3. The table *original_data* is table having the columns *the_geom* (feature geometry), *gid* (general primary key) and *class_id* (class_id of particular class that this feature belongs to. The column *class_id* is foreign key that references to the classification table called *original_classes*. The table *original_classes* contains classes' enumeration and has column *class_id* (general primary key) and column *class_desc* that contains textual description of such class.

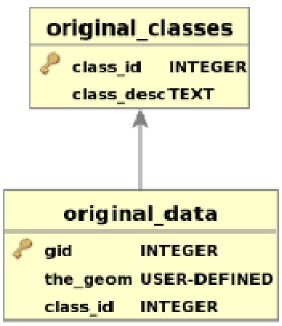

Fig. 3. Data model of the source data.

The graphical representation is shown on left part of fig. 4.

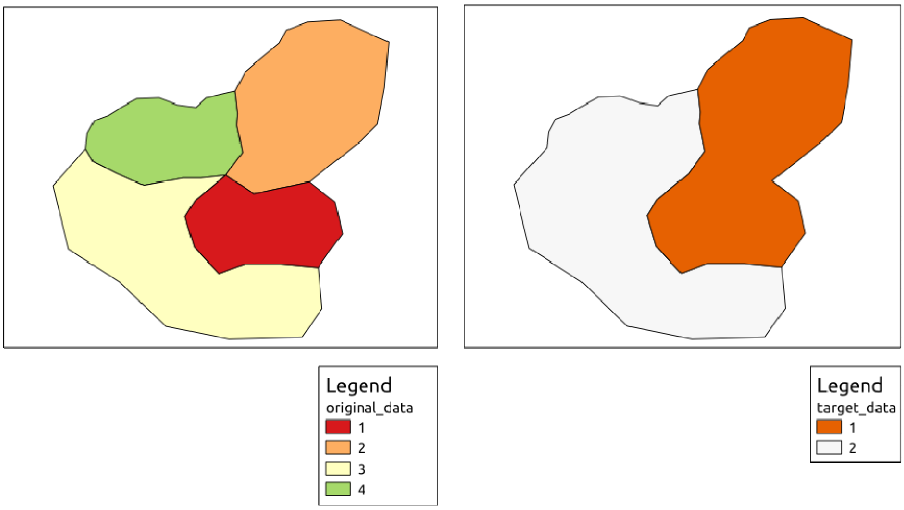

Fig. 4. The original data on the left side, harmonized data on the right side.

Let's assume that the target classification is simplified just to two classes (table 1).

Table 1. Target classification

class_id	**class_desc**
1	Forests
2	Open Space with little or no vegetation

According to knowledge from data experts we can put together the mapping between source and target classes. This mapping table is shown in table 2.

Table 2. Mapping between the original and target data

original_class_id	**target_class_id**
1	1
2	1
3	2
4	2

When using foreign keys to ensure data consistency we get schema shown at figure 5.

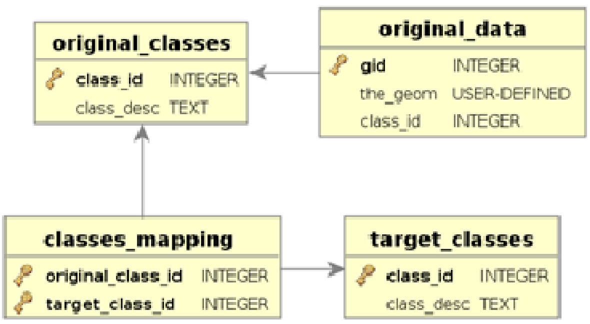

Fig. 5. Database schema using foreign keys for ensuring the data consistency.

Now we can use SQL JOIN to get reclassified result that is shown in table 3.

Table 3 Result of SQL reclassification query.

gid	the_geom	original_class_id	original_desc	target_class_id	target_class_desc
4	POLYGON((6.957...	4	Sparsely vegetated areas	2	Open Space with little or no vegetation
2	POLYGON((10.87...	2	Coniferous forest	1	Forests
3	POLYGON((6.359...	3	Bare rocks	2	Open Space with little or no vegetation
1	POLYGON((10.03...	1	Broad-leaved forest	1	Forests

Last step is the union of adjacent geometries. This can be done by aggregation with spatial union function.

The visualization of target data is shown on right part of fig. 4. It is worth to mention that union function might be demanding for larger dataset and the particular implementation and proper usage of indexes can greatly influence that.

4. HARMONIZATION USING GIS

Against the previously presented example which was focused on taxonomies, tools presented in this chapter are more suitable for harmonization of geometries. The particular problem to solve in this example was following: there was designed INSPIRE compliant data structure for land cover [2] in the scope of the Plan4all project. This project aimed on contribution to the standardization in the field of spatial data from spatial planning point of view. The main aim of the project was to harmonize spatial planning data and related metadata according to the INSPIRE principles. To show the way how to harmonize national data models for land cover into the designed model, there was created an example of harmonization land cover data of Zemgale Planning Region (Latvia). This harmonization was realized using ESRI technologies.

According to the 5-step harmonization approach framework, there is of course necessary to understand both source and target data structure (scheme). Identification of used coordinate systems and matching relevant land cover categories was the key part in this particular case. The harmonization process can be shortly outlined by following steps:

- Conversion of coordinate system
- Transform the source data geometry and attributes to match the target scheme.
    - Load both source data and target scheme to a workspace.
    - Convert the source classification to the target classification.
    - Dissolve boundaries of neighboring areas with the same standard classification.
    - Import the data into target templates and create relationships.

Technical realization is based on ESRI technology: ArcGIS Desktop provides the complete framework, Geodatabase is used as a database-based format for data storage and ModelBuilder allows to chain atomic geoprocessing operations into complex process model, which solves above described task. Further description focuses primarily on process models in ModelBuilder environment. Please see [5] to understand the legend of the following model.

The first used model (see figure 6) contains a definition of transformation coefficients and transforms source data from original coordinate system (LKS92) into coordinate system of target data structure (WGS84).

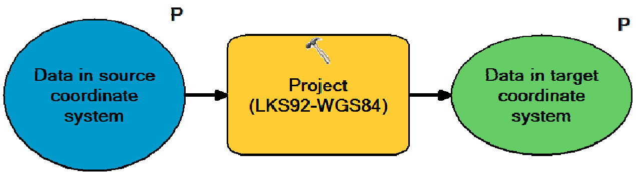

Fig. 6. Model for coordinate system transformation.

The reason, why the main model (due to its complexity not figured here) is so complex is caused by the nature of geoprocessing (each atomic operation needs input and output data), but there are two fundamental parts of the model: matching source attributes to target scheme and dissolving inner boundaries of polygons with identical target (standard) classification.

The next model (see figure 7) shows how it is possible to use an enumeration and apply it as a domain to a field. This model is not unnecessary for harmonization but it depicts how to create a domain, which can protect important field, particularly the *standardClassification* against incorrect input values.

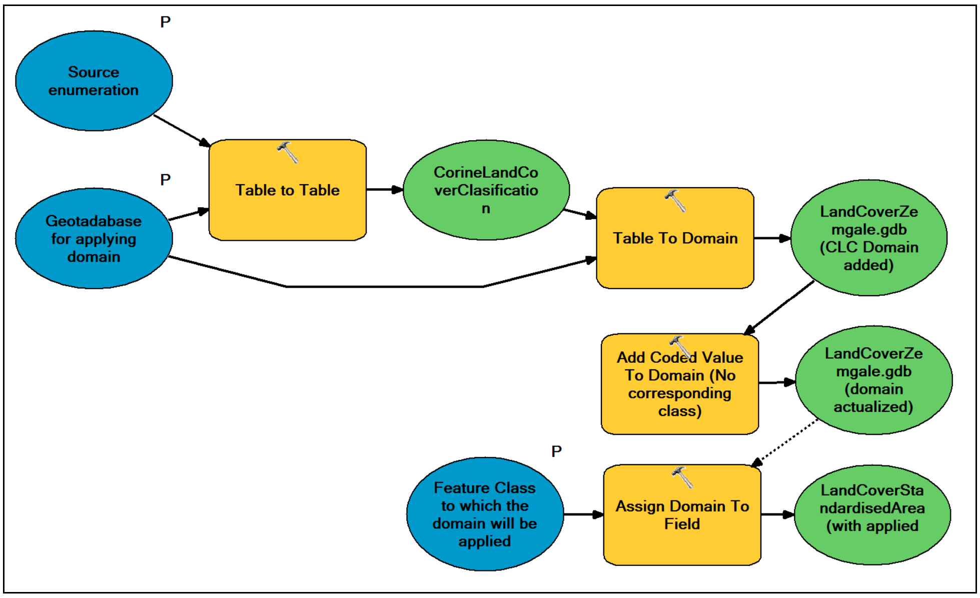

Fig. 7. Model for coordinate system transformation.

5. ONTOLOGIES AND ETL TOOLS

The case study is composed of two main technological parts – ontology (including a *reasoner* - software component making possible to derive new information based on existing, explicitly written information) and ETL (Extract-Transform-Load) tool. As the domain the spatial data of themes *land cover* and *land use* (as defined in Annex II and

Annex III of INSPIRE directive) was selected. The main barrier limited the combination and sharing of these data is cause by a huge number of classification systems (nomenclatures, hierarchies). This is the reason why the harmonization (mapping, transformation) of classification systems is the most required harmonization sub-process. The same situation is also in the other spatial data themes (e.g. Natural Risk Zones, Habitats and Biotopes etc.) that are based on some taxonomy.

The ontology is used to formalize second and third step of 5-steps harmonization framework (Source data understanding, Target data understanding). As it was mentioned in the previous paragraph the heterogeneity of spatial data of themes *land cover* and *land use* (LC/LU data) is caused by different classification systems. The ontologies enable to describe two or more taxonomies and their items using parameters characterized main properties (e.g. type of surface, use or vegetation). Source and target data coded in ontological systems are processed by the *reasoner*. This *reasoner* creates new inferred ontology containing combination of asserted (original) ontologies and so transformation rules between source and target data.

The developed transformation rules can be inserted to transformation part of ETL tool – practical realization of harmonization. This software makes possible to create *job* describing the complete harmonization (sub-) process, including open source data set(s), their processing (transformation) and saving new data. The complete case is shown on the Fig. 8.

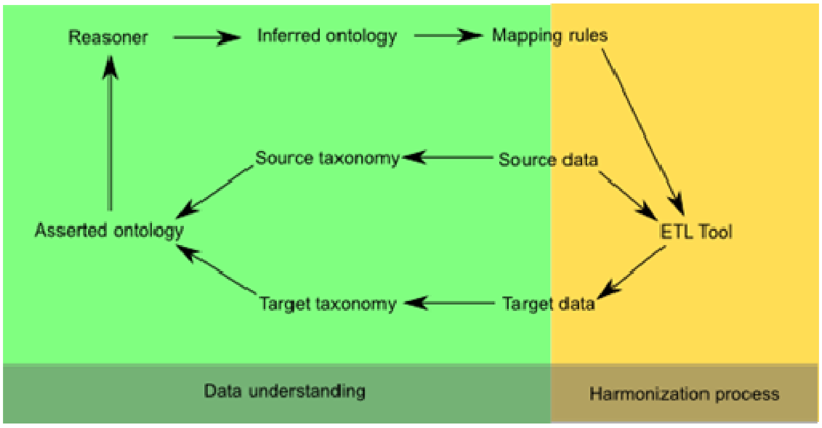

Fig. 8. Harmonization process using ontologies and ETL tool.

5.1 THEORETICAL OVERVIEW OF BASIC COMPONENTS

To realize above mentioned harmonization process two types of software were used – ontologies and ETL tool. One of the first definitions of ontologies in the information technologies was formulated by T. Gruber – "An ontology is an explicit specification of a conceptualization." [6]. More than ten years later he described the ontologies in the following way [7]: "In the context of computer and information sciences, ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. In the context of database systems, ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modelling knowledge about individuals, their attributes, and their relationships to other individuals. Due to their independence from lower level data models, ontologies are used for integrating heterogeneous databases, enabling interoperability among disparate systems, and specifying interfaces to independent, knowledge-based services."

W. Borst [1] also extent the original Gruber's definition: "An ontology is a formal specification of a shared conceptualization". R. Poli [10] recommended to use and distinction between two words connected with ontologies – formal (application of formal terms) and formalized (formal way of coding).

According to [3] for the harmonization purposes the domain ontology focused on terminology (classification systems) of LC/LU data was used. The publication [4] explains that "domain ontologies are intended to provide a source of predefined concepts that provide descriptions about specific domain knowledge".

Common ontologies (created in OWL /Web Ontology Language/ data format) contain five essential elements or groups of elements (class, relation, instance, axiom, annotation). In our transformation ontology mainly used classes and their relations (object and data types).

ETL tools are not based on the large theory as ontologies. These tools ensure complete management of data. It means possibilities of loading and saving data sets in different formats, including databases. The uploaded data can process in many ways (e.g. statistics, database operation, spatial data analyses etc.). There are two main advantages of the ETL tools:

- graphical design of data processing,
- using just one software.

5.2 HARMONIZATION PROCESS

The harmonization process is composed of two parts. The development of mapping rules (data understanding) is mainly provided by transformation ontology. The transformation ontology was developed and is managed in the software Protégé, version 4.1 (protege.stanford.edu) and written in the format OWL 2.0 (http://www.w3.org/TR/owl2-overview) using OWL/RDF (Resource Description Framework) syntax. Transformation ontology is complex ontological systems divided into three parts (Fig. 9):

- File Parameter.owl containing 9 parameters of the classes of particular classification systems (Climate, Formation, Homogenity_and_continuity, Nature, Season_of_crop, Surface, Use, Vegetation and Water_management) and allowed values of these parameters (e.g. parameter Climate has three values – Mild, Subtropical, Tropical).
- Files in the middle of the Fig. 4 contain all classes of particular classification systems. These classes are described by parameters and their values (e.g. fragment of description of CLC class 5 Water bodies – *...hasVegetation some AnyVegetation and hasSurface some Water...*).
- File Transformation.owl interconnects all above-mentioned files. The mapping rules between taxonomies are developed by processing of this file using reasoner (integrated to ontology editor). The new inferred ontology combines all classification systems and show relations between their items.

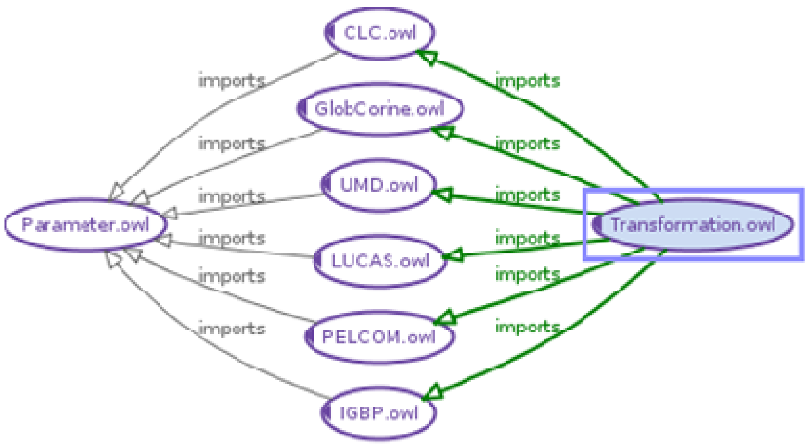

Fig. 9. Scheme of the transformation ontology files.

The ETL tools arrange for the second part – concrete harmonization. There are two major open-source ETL applications that enable spatial data processing – GeoKettle and Talend Open Studio with integrated Spatial Data Integrator (www.talend.com). The LC/LU harmonization case was realized in the second one software package. The harmonization could be realized according to the scheme (Fig. 10). There are four parts (described from the left side):

- Uploading of source data (in ESRI Shapefile format) to ETL tool.
- Component tMap_2 enable to modifying data models of source and target data (for instance changing attributes, their names or types).
- Third part is used to re-classification. Based on transformation ontology (inferred version of this ontology) the mapping table was created.
- Storing of target data (also in ESRI Shapefile format, but Open Studio supports other formats including databases).

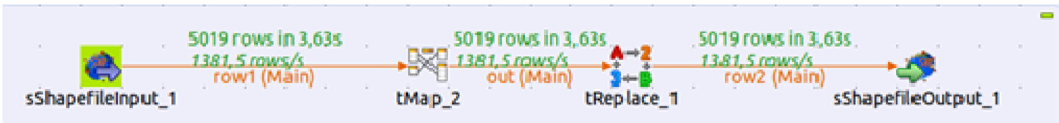

Fig. 10. Harmonization processing line.

6. CONCLUSIONS

The table 4 shows the main pros and cons of proposed harmonization solution based on ontologies and ETL.

Table 4 Ontologies & ETL tools - pros & cons.

Advantage	**Disadvantage**
All used software tools (Protégé, reasoners, Talend Open Studio and Spatial Data Integrator) are free of charge and mostly open-source.	Java language is not fitting programming language to develop an application processing a large number of data (the processing is slow).
All used software is developed in Java language therefore the complete solution is	The proposed solution was not tested together with large data set.

Advantage	Disadvantage
multiplatform and independent on operational systems. The similar advantage follows from XML based data formats (like OWL).	
The whole proposed solution is modular and open to extension and modification.	It is necessary to understand to principles of ontologies.
Both main parts offer intelligible graphic view enabling design, building and checking of harmonization subprocesses.	
ETL tools (in general) are able to combine spatial and non-spatial data.	
Ontologies can interconnect the proposed systems for LC/LU classifications description with similar existing solutions.	

From the relational database management systems point of view these systems are definitely one of the software products that have proven their concept in the past. They have very clear and unified query model that is a result of long research work and this model is also fully adopted in common knowledge of IT engineers.

Once it is possible to import our data to RDBMS we can utilize existing functions and SQL language to fulfill harmonization processes focused on change of taxonomies and any related harmonization of attributes (rename, retype) as well as some of the geometry processing (depending on available spatial functions of particular RDBMS. When using RDBMS it is also very valuable that we can leverage the properties of query planning and spatial indexes that can greatly optimize whole process.

Using the commercial software for 5-steps harmonization framework, there can be set some generalizable conclusions from this particular example:

- Model is very clear tool but even simple operations require a long model, whose sometimes are not easy to read.
- This technology is suitable for organizations whose already works with ESRI platform, but it is a quite expensive solution – and there exists open source alternatives.

The 5-steps harmonization framework is a robust solution and mechanism how to deal with data harmonization. Practically it was demonstrated with real spatial planning data. As the trend is to share and provide data in spatial data infrastructures on various levels (local, national, European) the importance of data harmonization will be still increasing.

ACKNOWLEDGMENTS

The research was supported financially by the project Plan4all - European Network of Best Practices for Interoperability of Spatial Planning Information (registration number ECP 2008 GEO 318007) and by the European Regional Development Fund (ERDF), project “NTIS – New Technologies for the Information Society”, European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

REFERENCES

[1] Borst, W. N. (1997). Construction of Engineering Ontologies for Knowledge Sharing

and Reuse. University of Twente.
[cit. 6. 10. 2011]. URL: http://doc.utwente.nl/17864/1/t0000004.pdf.

[2] Cerba et all. (2012) Conceptual Data Models for Selected Themes. Technical Report. Plan4all - European Network of Best Practices for Interoperability of Spatial Planning Information. [cit. 9.10.2011] URL: http://www.plan4all.eu

[3] Cerba, O. (2011). Ontologies in Cartography: Power of Reasoning. In Joint ICA Symposium, Orleans (France).

[4] Dolbear, C. et al. (2005). Semantic interoperability between topographic data and a flood defence ontology. Technical Report. Ordnance Survey. [cit. 6. 10. 2011]. URL: http://www.ordnancesurvey.co.uk/partnerships/research/publications/docs/2005/Semantic%20Interoperability_CDolbear_sem.pdf.

[5] Environmental Systems Research Institute, Inc. (2010). Model elements. ArcGIS Resource Center. [cit.4. 10. 2011].
URL: http://help.arcgis.com/en/arcgisdesktop/10.0/help/#/Model_elements/002w00000003000000/

[6] Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199-220.

[7] Gruber, T. (2009). Ontology. *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Ozsu (Eds.), Springer-Verlag. [cit. 4. 10. 2011].
URL: http://tomgruber.org/writing/ontology-definition-2007.htm.

[8] INSPIRE Drafting Team Data Specifications (2009). D2.5: Generic Conceptual Model. [cit.12. 8. 2011].
URL: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/D2.5_v3.2.pdf

[9] Jezek, J., Cerba, O. (2011) Technological aspects of spatial data harmonization. *Proceedings of Sympozium GIS Ostrava 2011*. Ostrava. ISBN: 978-80-248-2366-9.

[10] Poli, R. (2002) Ontological methodology. *International Journal Human-Computer Studies*, 56, s. 639-664. Elsevier Science Ltd.

[11] The European Parliament and the Council of the European Union (2007). Infrastructure for Spatial Information in the European Community (INSPIRE) Directive. [cit. 10. 7. 2011]. URL: http://inspire.jrc.ec.europa.eu/