HUMBOLDT ALIGNMENT EDITOR

Jiří HORÁK[1], Lucie JUŘIKOVSKÁ[2] , Miroslav UMLAUF[3] , Jan JEŽEK[4] , Markéta HANZLOVÁ[5], Sisi ZLATANOVA[6]

[1,2,3] Institute of Geoinformatics, HGF, VSB-Technical University of Ostrava, 17.listopadu 15, 70833, Ostrava, Czech Republic
jiri.horak@vsb.cz, lucie.jurikovska@vsb.cz, miroslav.umlauf@vsb.cz

[4] Section of Geomatics, Faculty of Applied Sciences, The University of West Bohemia, Univerzitní 8, 30614, Plzeň, Czech Republic
jezekjan@kma.zcu.cz

[5]GISIG - Geographical Information Systems International Group, Via Piacenza 56, 161 38, Genova, Italy
m.hanzlova@gisig.it

[6]Delft University of Technology, OTB Research Institute, Jaffalaan 9, Build. 30, 2628 BX Delft, Netherlands
S.Zlatanova@tudelft.nl

Abstract

The process of data harmonization starts with a description of data sources and mapping of differences among heterogeneous source and target data sets.

The semantic differences usually include incompatible data schemas (structures, different data types, domains etc.) and differences in content (different values for the same entity). While the second type of differences is difficult to be solved in an automated way, the first type can be described by the new tool HALE, developed within the HUMBOLDT project. The HUMBOLDT Alignment Editor (HALE) allows to interactively specify transformations between models at a conceptual schema level. It is possible to rename attributes, to make reclassification (classification mapping), to change geometric types, to change values using various mathematical expressions, to set default values etc.

The paper presents features, functions and results of testing harmonisation process for selected geodata. The process of data integration into common schemas which are INSPIRE compliant is discussed. The INSPIRE theme Hydrography is used as a core of the solution. Finally the schema transformation performed with HALE is executed with the help of other HUMBOLDT tools (i.e. Conceptual Schema Transformer) to perform required data transformation.

Keywords: Humboldt, harmonisation, data structure, alignment, conflation, WFS, GML, INSPIRE

INTRODUCTION

The purpose of data harmonisation is to resolve heterogeneity in spatial data and make them interoperable. This means that the systems have to be able to provide tools to combine seamlessly all available data. Traditionally three types of data interoperability are identified namely system interoperability, syntax and structure interoperability, and semantic interoperability (Sheth, 1999).

System interoperability reflects operating systems and communications heterogeneity, e.g. the instruction sets, communication protocols, different file systems, naming, file types, operation and so on. As a part of system interoperability it is possible to specify syntactic interoperability. Bishr (1998) and Fonseca et al. (2000) describe syntactic heterogeneity, in which the databases use different paradigms. Stuckenschmidt (2003) explains syntactic heterogeneity for GIS applications using differences in data format.

Structure (schematic) interoperability refers as to data models, data structures, data domains and data representation. Bishr (1998) specifies schematic heterogeneity, in which the same object of the real world is represented using different concepts in a database.

Semantic interoperability is the most complex one and deals with the meaning of the data. Semantics refers to the aspects of meanings that are expressed in a language, code, message or any other form of representation, i.e. semantic interoperability requires that the information system understands the semantics of the users' request and information sources (Sheth 1999). Bishr (1998) and Fonseca et al. (2000) explain semantic heterogeneity, in which a fact can have more than one description. Stuckenschmidt (2003) emphasizes differences in intended meaning of terms within specific context for this type of heterogeneity. Semantic heterogeneity should be solved before schematic and syntactic heterogeneity (Fonseca et al., 2000) using semantic translators (mediators).

It is worth to mention that in the case of datasets overlays a problem of content heterogeneity (different values for the same fact) has to be solved.

Much research has been completed through the years on all aspect of heterogeneity by researchers and standardization organisations. Open Geospatial Consortium (OGC) and ISO (International Standards Organisations) are currently working on standards to solve syntax heterogeneity. Spatial Data Infrastructures (SDI) are being built by different regions, countries and even across national borders (Bernard 2002, Groot and McLaughin 2000, Riecken et al. 2003), are examples of resolving syntax heterogeneity. SDIs support the discovery and retrieval of distributed geospatial data sources and can provide tools to resolve syntax interoperability but only to certain extends (Lutz and Klien 2006). Harmonisation processes represent an important, core part of building SDI. Methods of harmonisation data, metadata, processes, functions, procedures and rules are essential for creating consistent and operational SDI where end-users may access and employ in their systems different data stored in different places in different structures using different rules.

The HUMBOLDT project contributes to solving structural and semantic interoperability and is specifically focussed on the implementation of an European Spatial Data Infrastructure (ESDI). It provides tools, which allow for integration of spatial data sets from the multitude of European organizations. It was the aim of this project to manage and to advance the implementation process of this ESDI. The HUMBOLDT design approach re-uses existing concepts, processes, implementations and experiences as discussed in research articles and standardisation documents. The most important HUMBOLDT tool that deals with structural and semantic heterogeneity is the HUMBOLDT Alignment Editor (HALE). HALE is a desktop application that allows to design data transformations interactively at a conceptual schema level. This paper presents our tests with the HALE tool performed on one of the HUMBOLDT scenarios, i.e. Transboundary catchment.

The paper is organized in the following order: next section elaborates further on SDI and data harmonization issues. Further section presents the overall HUMBOLDT framework for data harmonization. Section HUMBOLDT Alignment Editor presents and discusses the schema mapping tool HALE. Section Case study: Transboundary catchment Roya/Roia river elaborates on the scenario Transboundary catchment and the tests with HALE. The final Section discuss the results.

SDI AND SPATIAL DATA HARMONISATION

Spatial Data Infrastructure (SDI) is the core of any GeoInformation Infrastructure (GII). SDI enables to integrate different spatial data sources and build seamless databases and data-portals providing a central place how to access different data. Roles of metadata, standardisation and geodata sharing (using web services) are essential for SDI design (Nougeras et al., 2005). Ideally SDI should provide means which would hide original data structures, formats and places of storage, and it should offer a transparent access to spatial data (no matter of original way of data storage).

The current implementation of European SDI is closely linked with European initiatives like INSPIRE, GMES and SEIS. The National Geoinformation Infrastructure of the Czech Republic (NGII) has been prepared since the end of 1990, supported by CAGI and Nemoforum (Národní geoinformační infrastruktura České republiky, 2001); nevertheless a significant acceleration of the real SDI establishment was connected with launching the INSPIRE directive (Infrastructure for Spatial Information in Europe, 2007/2/EC). INSPIRE declares

necessity to collect data once and maintained it at the level where most effective; and to combine spatial information from different sources seamless and shared between users/applications (Pauknerová, Tryhubová 2006).

Successful implementation of INSPIRE is conditioned by shared data and services compliant to common standards. Main data specifications for INSPIRE can be found on the data specification page (http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2) on the INSPIRE website. Among various documents important for design of INSPIRE compliant systems, it may be worth to note following documents:

- INSPIRE Data Specification on individual domains (e.g. Protected Sites, Hydrography)
- INSPIRE GML Applications Schemas (http://inspire.jrc.ec.europa.eu/index.cfm/pageid/541/downloadid/1707) (available in the "Other Documents" section).

The large requirements for geodata harmonisation lead to extended activities tied up with ways of facilitating the harmonisation process by automating the necessary steps as far as possible. Table 1 provides and overview on some of the harmonisation aspects and the possible implementations, i.e. either off-line of online.

Table 1 Review of harmonisation requirements (Vries, de, M. et al, 2010)

Harmonisation goal or purpose	***Offline/pre-publishing (preparation)***	***Online/during use (runtime)***
data (exchange) format	Conversion tools such as FME (Safe Software), export modules of GIS/CAD software, all kinds of image processing and conversion tools	Web services with standardized interfaces that act as 'wrapper' around native formats and produce standard formats (raster of vector) as output (e.g. WMS, WFS, WCS, Web3D)
spatial reference system, reference grids	Beforehand, e.g. have a copy in WGS84 or ETRS89 in case fast retrieval is important	Coordinate transformation by web server of data provider, or by Web processing service, or in client
data/conceptual model: structure and constraints	1. defining common model and constraints (UML, OCL, OWL) 2. establishing transformation rules from local to common model (INTERLIS) 3. encoding transformation rules in machine-readable format (sql, XSLT, OWL, QVT/ATL) 4. migration or replication	a. Transform to target model by Web service of data provide or by cascading Web service (WFS-X, mdWFS), b. Or mediate to target model by separate mediator Web service(s), c. or translate to/from target model by client-side software
nomenclature, classification, taxonomy	Defining common nomenclature and classification or taxonomy	Use the standardized classification/taxonomy in: metadata, in search engine (keyword lists), in data content (code lists), for generalization (offline or real-time)
terminology/vocabulary, ontology	Terminology and definitions in thesaurus, data dictionary and/or ontology	
metadata model	Define a ISO 19115 or Dublin Core profile and migrate metadata to that common metadata model	Either centralize the metadata registry, or have distributed registry nodes
scale, amount of detail, aggregation for reporting	MRDB or vario-scale databases, for thematic aggregation: taxonomies and/or ontologies	(cartographic) generalization/refinement in real-time

Harmonisation goal or purpose	*Offline/pre-publishing (preparation)*	*Online/during use (runtime)*
Portrayal (legend/classification, style)	1. defining standards, e.g. IHO S52, DGIWIG 2. encoding in machine-readable format	Applying rules, e.g. by using SLD or default styling in GML/WFS
Processing functions: their parameters and formulas/algorithms	Agreement on parameters etc. and describe in repositories (possibly same as data and service metadata registries)	e.g. Web processing services that retrieve functions and parameters from repositories
extension (spatial, thematic, temporal)	Data quality actions like edge matching. But also detection of doubles (solve conflation issues)	
data collection procedures	e.g. guidelines for digitizing	

The HUMBOLDT project addresses many of above mentioned issues and concentrates on development of appropriate tools to support automated harmonisation processes. A special attention is dedicated to the usage of web based tools aiming to create an open and distributed system easily integrated to various portals as well as an individual application. Explanations of standard geoweb services can be found in (Lehto, L. and T. Sarjakoski, 2004, Charvát et al., 2007, Šeliga et Růžička, 2005). The interoperability of geoweb services is addressed by several projects; capabilities of semantic oriented geoweb services are introduced in (Vaccari et al., 2009).

HUMBOLDT FRAMEWORK

A core development within the Humboldt project is the framework for data harmonization. The framework stands for a set of software tools and libraries that helps other developers or advanced GIS users to set up infrastructure for data harmonization. The main concept is described in figure 1. Humboldt tools are developed in Java programming language and licensed under LGPL (open source license). Humboldt framework consists of desktop application (e.g. HALE), software libraries (e.g. CST) and server-side application (e.g. CST-WPS, Edge Matching Service). From the technical point of view, Humboldt components are standalone modules that are based on Maven build system. Most of the components can be also used as OSGi bundles. The development of Humboldt framework is based on existing ISO and OGC standards and influenced by other projects like CUAHSI (http://www.cuahsi.org/).

The general schema of data harmonisation is depicted in figure 2 using a data flow diagram. As it can be realised, two basic phases can be distinguished. First, harmonisation steps have to be designed using one of the mapping tools e.g. HALE, WDCS. The next phase solves the actual data transformation (i.e. the transformation of the data sets) according to the harmonisation schema. Data harmonisation implementation utilises other HUMBOLDT tools or other suitable tools.

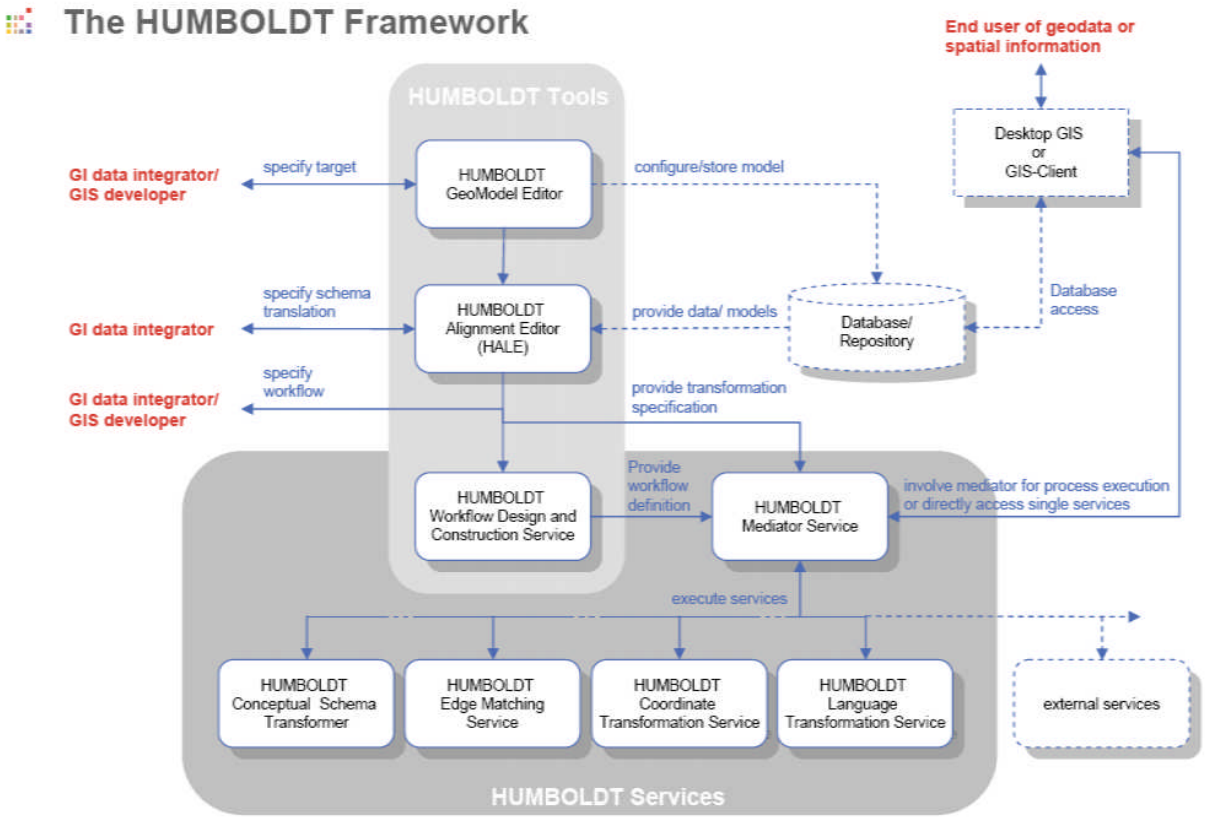

Fig1 Concept of HUMBOLDT Framework

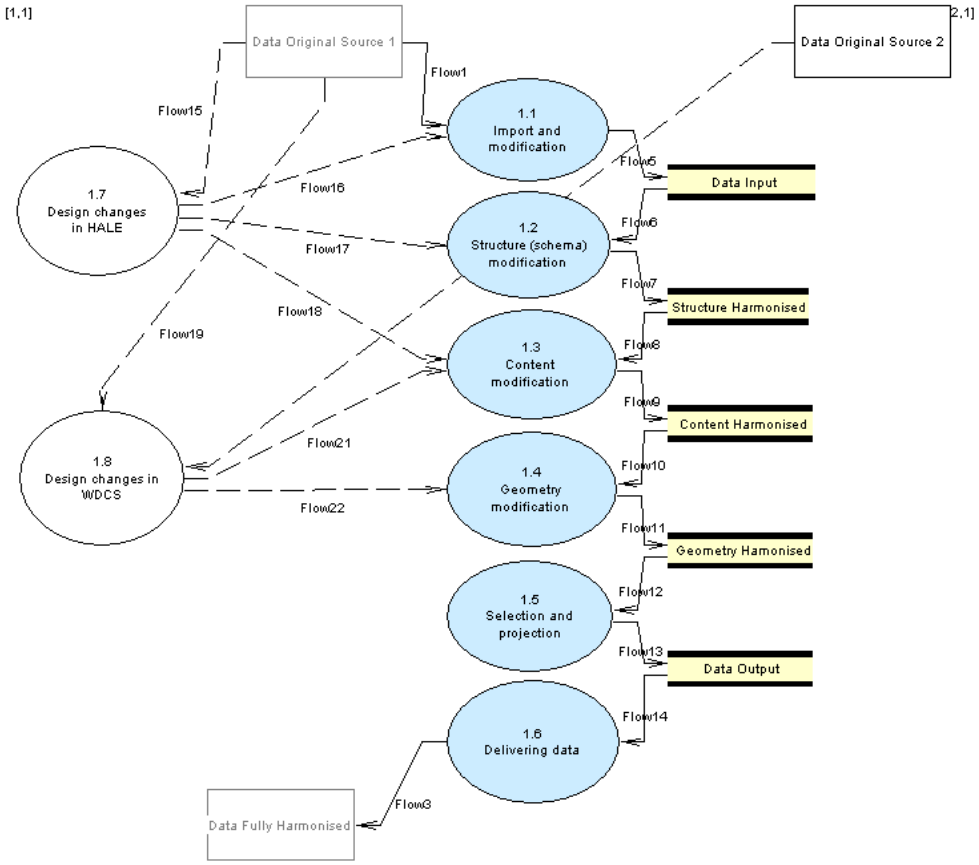

Fig2 Data Flow Diagram of the harmonisation process for the data source 1

It is worth to mention that not all of HUMBOLDT components reached the stable phase of development, but the main goal to prove the concept and establish overall API was achieved. One of the most developed products are HALE (the schema mapper) and CST (the executor of the schema mapping). The status of all components can be tracked on Humboldt community site (http://community.esdi-humboldt.eu/).

The main blockers to reach the stable status are:

- Handling of GML 3.2.1 – this version of GML is required by INSPIRE but has not been yet widely adopted by other GIS software libraries and application. The support of such encodings is nowadays still limited so even if Humboldt framework can generate such outputs there is not many possibilities to use it in other third party software.
- Handling of GML in general – even if GML is a OGC standard, its implementation by third party vendors is not consistent. GML output of different software products (ogr2ogr, Geoserver, Geomedia) has always its specifics. One of the reasons might be high level of complexity of GML (Galdos, 2010).

HUMBOLDT ALIGNMENT EDITOR

The HUMBOLDT Alignment Editor (HALE) is an Eclipse RCP application (Reitz 2010a, Reitz and Kuijper, 2009. HALE allows to interactively define mappings between source and target conceptual schemas. It supports import of schemas (i.e. Eclipse ecore, WFS and GML Application Schemas) and provides tools to indicate mappings between classes, their attributes and relations. Several different cases can be distinguished while mapping classes, relationships and their attributes (Lehto 2007, Reitz 2010b). HALE performs a large number of the specified mappings (see below). As discussed in the HUMBOLDT framework, the defined transformations are stored either locally or in the HUMBOLDT Model Repository and used by the Conceptual Schema Transformer (CST) to perform Schema Transformation on actual geodata. The HALE installer is available on http://community.esdi-humboldt.eu/projects/list_files/hale.

Features

HALE allows resolving several interoperability issues:

- Differences in application schema and terminology. HALE provides mapping rules for the classes and attributes of a source to a target conceptual schema.
- Differences in Metadata. HALE is able to create mapping rules for metadata elements.
- Inconsistency in selected spatial and temporal aspects. HALE enable to define functions for transformation of geometric types.
- Multiple representations. HALE will offer a definition of rules for handling other representations of the same object, i.e. under what circumstances which of the precedence should be used.

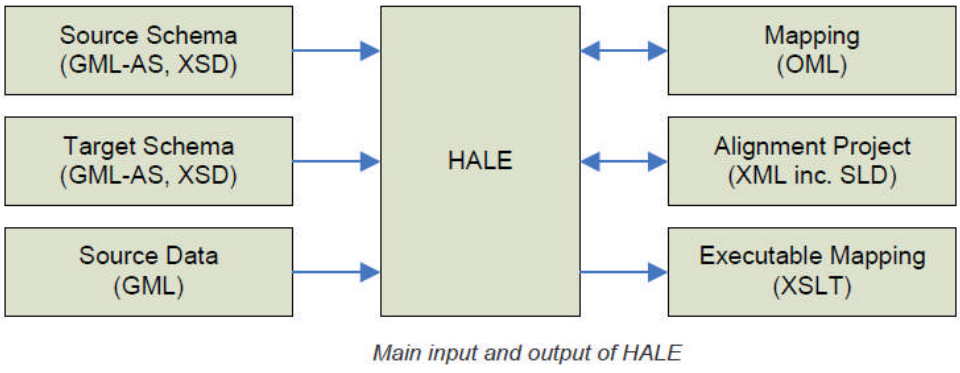

Fig3 Main inputs and outputs of HALE (Reitz, 2010)

The list of proposed functions includes:

1. A simple Attribute Rename Function. It facilitates the change of alphanumeric attributes and also mapping the following geometric types: LineString -- MultiPoint, Polygon -- MultiPoint,
2. A Generic Math Function - calculation of mathematical expressions using source attributes as variables,
3. A Bounding Box Function and a Centroid Function – creation of additional geometry (polygon - minimum bounding rectangle MBR; centroid - point),
4. A Buffer Function – creation of buffer (polygon) around any line or point-type geometry,
5. A Classification Mapping Function – transformation of code lists and classification values.
6. An Ordinates to Point Geometry function that accepts two numerical values as input and will create a point geometry based on that.
7. A Date Extraction Function that allows to extract a date object by specifying the format of that date and to set it in another format.
8. A Create GML Reference Function enables to create a reference to another element in the same data set or a different data set.
9. An INSPIRE Identifier Function enables to create a IdentifierProperty-Type;
10. An INSPIRE Geographic Name Function does the same for Geographic-NamePropertyTypes.

It is possible to classify functions into following main categories: Create new spatial objects, Structure modification and Content modification (Table 2).

Table 2 Classification of HALE functions

Category	Function	HALE function
Create new spatial objects	identification of key attribute	INSPIRE Identifier Function
	geographic name	INSPIRE Geographic name
	MBR	Bounding box function
	point from text	Ordinates to Point Geometry
	point – centroid	centroid function
	buffer	buffer function
Structure modification	attribute name change	rename attribute
	geometry datatypes (i.e. polygon to linestring)	rename attribute
	Integrity constraints change (i.e. adding PK, unique identifiers, null check, referential integrity, user defined IC)	create GML reference function
Content modification	Fill by a given value	Attribute default value
	Fill by NULL value	INSPIRE Nil reason
	Fill by a numerical expression	mathematical expression
	Replace a date (change format)	date extraction
	Replace strings	classification mapping

Not all functions are available in the current version of HALE.

HALE produces three types of outputs:

- GOML files contain information for mapping. It represents an input for HUMBOLDT Conceptual Schema Transformation Service and manages the data scheme transformation.
- XML file containing a configuration for HALE project (used only by HALE).
- SLD file intended for geodata visualisation (used only by HALE).

The mapping can be also saved in OML or XSLT format.

CASE STUDY: TRANSBOUNDARY CATCHMENT ROYA/ROIA RIVER

Water management and hydrological modelling

Water management in transboundary water catchments strongly relies on collaboration of stakeholders from both sides of the border. Any integrated water management requires a joint effort and data interchange to reach adequate decision support.

Hydrological modelling provides various tools which may be successfully exploited in water management.

Principles of numerical modelling of hydrological processes and description of commonly used methods can be found in Bedient and Huber (2001), Maidment (1993) and Beven (2002). A practical evaluation of 18 most frequent numerical modelling systems for water management was provided within the framework of the TANDEM project. The following features were investigated (Horák et al., 2008): embedded models, field of applications, interoperability (linkage to GIS, utilisation of Earth coordinate systems, remote management and control like a macro language or an application programming interface), price and license terms, support (updating, technical support, documentation), software features (operating system, modularity, user interface, provided functions, possibility of integration), input and output (obligatory, conditional, optional). The type of modelling and type of software implementation determine data requirements for numerical hydrological modelling.

Data required for hydrological modelling

A list of data required for hydrological modelling includes hydrometeorological data (mainly time series of rainfall data, records of river discharge) and geographical data necessary for setting of conditions influencing hydrological processes (i.e. transformation rainfall into a water flow).

Table 3 Main data requirements for hydrological modelling

	Required attributes	**Description**
Digital Elevation Model	Altitude	3D digital representation of the topography
River Network	Width	
Water bodies	Type Altitude	Included Lakes, Reservoirs, etc.
Catchment Area		Catchment boundaries of interest, watershed geometry.
Land Use, Land Cover	Code/CLC Code Vegetation Type	Land cover/Land use data, and other vegetation data (LAI, ...)
Soil	Type	soil data e.g. layer depth, bulk density, porosity, field capacity, saturated conductivity) intended to derive some hydraulic and runoff coefficients

Roya/Roia river catchment

The Roya/Roia River basin is one of the internationally shared river basins which crosses France and Italy. The Roya/Roia River catchment covers area about 675 square kilometres. The river springs in an altitude of 1871 m above sea level at Col de Tende on the French-Italian border and runs south about 35 km to Breil-sur-Roya and then another 8 km where it crosses back into Italy and discharges directly to the Mediterranean Sea in Ventimiglia, with an average flow of 15 cubic meters per second. The morphology of the basin is significantly different in France and in Italy. The landscape in France (Upper Roya) is the larger part of the contributing area and is characterized by mountains and valleys with swift rivers, thick forests, and 'Italian-wise' villages. In the southern part, the Italian territory (Lower Roia), the Roia flows in a flood plain area.

Hydrological modelling is required for understanding hydrological processes in the catchment, designing appropriate measures and improving water management (e.g. flash floods, occurring in the surrounding area of Breil Sur Roya, Figure 4).

Fig 4: Roya-Roia and danger of flash floods (Lac de Meshes,Breil sur Roya, Ventimiglia)

As the history shown, Roya basin is exposed for flood hazard and land slide. Some of the significant flood hazards and its consequences happened in the past and giving the corresponding time they are documented by Mitiku (2009).

Data from adjacent (collaborating) countries

Geographical data available from the both countries include contour lines, river network, water bodies, land-use and coast line. The French data were obtained from BD CARTHAGE® database provided by French National Institute of Geography, IGN (Institut Géographique National). The Italian data were provided by Regione Liguria. The French and Italian data obtained were converted into GML files and datasets were made available on Geoserver provided by GISIG to the scenario working group (http://www.gisig.it:8081/geoserver). The French data were obtained in two coordinate reference systems (CRS), recognized under the name “NTF (Paris) / Lambert zone II” (defined by EPSG code 27572) and “RGF93 / Lambert-93” (defined by EPSG code 2154). The Italian CRS is “Monte Mario / Italy zone 1” (defined by EPSG code 3003).

Let us assume requirements for transboundary hydrological modelling originating from the Italian side. It is necessary to transform all French data from their CRS (EPSG code 2154 and 27572) to the Italian coordinate system (EPSG code 3003). Next, the layers with the same CRS have to be matched on borders and joined together. A horizontal conflation (Blasby et al. 2004) is needed for the following layers: contour lines, river network and land-use. This function is available in HUMBOLDT framework under Coverage Alignment (HUMBOLDT Edge Matching Service, see Fig.1). Finally, it is necessary to make transformation of the attribute structure and the attribute content. As described later, there are two basic possibilities of transformation - a one-side transformation from French data to Italian data (and vice versa) or more universal transformation to INSPIRE schema (reference later as two-side transformation) which means data from both countries is transformed into the INSPIRE compliant schema to make them joinable.

Two types of transformation processes were used: renaming attributes and classification mapping function. Only attributes and features necessary for hydrological modelling are adapted in the harmonisation process. The review of the number of required transformation for each layer into the target INSPIRE compliant schema is shown in table 4.

Table 4 Review of harmonisation requirements for hydrological data in the Roia/Roya catchment

Data	Rename attribute	Classification Mapping
Contour lines	2	-
River Network	4	3
Water Bodies	2(FR), 3(IT)	-
Land-use	3(FR), 2(IT)	2
Coast Line	2	-

Elevation

The digital elevation models are derived from contour lines in both countries. The vertical reference system is Genova 1942 (Italian data, Fig. 5) and NGF-IGN69 (EUREF, 2010) for French data (Fig. 6) and as defined in (IGN, 2002) IGN 1969 pour la France continentale and IGN 1978 pour la Corse. The important information given by this type of dataset is ALTITUDE that represents a mandatory attribute from the hydrological modelling point of view.

Property	Type
the_geom	MultiLineString
tipo	String
quota	Double
id	Long
origine	String
GDO_GEOMET	String

Fig 5: Italian Contour Line dataset preview with listed feature type details

Property	Type
the_geom	MultiLineString
FNODE_	Integer
TNODE_	Integer
LPOLY_	Integer
RPOLY_	Integer
LENGTH	Double
DEP_06_	Integer
DEP_06_ID	Integer
ALTITUDE	String
Object_ID	Integer
Shape_Leng	Double

Fig 6: French Contour Line dataset preview with listed feature type details

River and Water Body

Main differences between the different regions include different classifications of river width and different classifications of watercourse hierarchy. As to the water body layers, there are no polygons, which overlap. The important information given by this type of dataset is WIDTH that represents a recommended attribute from the hydrological modelling point of view. Although, Italian dataset does not provide such information and French dataset classifies such information as follows:

- "1" (from 0 to 15 m),
- "2" (from 15 m to 50 m),
- "3" (more than 50 m).

Land cover (CORINE)

Both datasets are classified according to CORINE nomenclature (3rd level CORINE code) with corresponding land cover description in national language. Example of different attribute names and values: first table is Corine land cover for France, the attribute called NIV3_06 has the same meaning as attribute CLASSE in the second italian table. The values in these two columns are also similar, but use different formats (Fig. 7).

NIV3_06	IDPOL_06	INTIT_06
111	1	Tissu urbain continu
111	2	Tissu urbain continu
111	3	Tissu urbain continu
111	4	Tissu urbain continu
111	5	Tissu urbain continu
111	6	Tissu urbain continu
111	7	Tissu urbain continu
111	8	Tissu urbain continu
111	9	Tissu urbain continu
111	10	Tissu urbain continu
111	11	Tissu urbain continu
111	12	Tissu urbain continu
111	13	Tissu urbain continu
111	14	Tissu urbain continu
111	15	Tissu urbain continu
111	16	Tissu urbain continu
111	17	Tissu urbain continu
111	18	Tissu urbain continu
111	19	Tissu urbain continu

CODICE	CLASSE	DESCRIZION
29	3.2.4	vegetazione bosc. ed arbust in evoluzione
32	3.3.3	vegetazione rada
25	3.1.3	boschi misti
29	3.2.4	vegetazione bosc. ed arbust in evoluzione
32	3.3.3	vegetazione rada
17	2.2.3	oliveti
26	3.2.1	aree a pascolo e praterie naturali
26	3.2.1	aree a pascolo e praterie naturali
25	3.1.3	boschi misti
25	3.1.3	boschi misti
2	1.1.2	tessuto urbano discontinuo
29	3.2.4	vegetazione bosc. ed arbust in evoluzione
29	3.2.4	vegetazione bosc. ed arbust in evoluzione
26	3.2.1	aree a pascolo e praterie naturali
26	3.2.1	aree a pascolo e praterie naturali
29	3.2.4	vegetazione bosc. ed arbust in evoluzione
26	3.2.1	aree a pascolo e praterie naturali
26	3.2.1	aree a pascolo e praterie naturali
21	2.4.3	aree con colture e spazi nat.

Fig 7: Description of the content transformation between French CLC data (let) and Italian CLC data (right)

Required data harmonisation for the Roia/Roya catchment

The top priorities of the harmonisation steps for transboundary catchments are:

- schema transformation, including the Classification Mapping,
- coordinate reference systems transformation,
- layers horizontal conflation (alignment).

Two basic types of transformation have been prepared:

- **one-side transformation**. Transformation of data source from foreign country to match own datasets and append data from the foreign dataset. Here, Italy is assumed to be the home country requiring hydrological modelling due to possessing lower part of the catchment. Thus the harmonisation process "French data ➔ Italian data" is demonstrated (table 5).
- **two-side transformation**. Data from both countries are transformed into the common target schema, which is typically INSPIRE compliant or INSPIRE based. Hereafter such schema is labelled INSPIRE.

Data profiles from both sides of the border, together with common data profiles (INSPIRE inspired), are instrumental for the target schema creation.

Table 5 Example of one-side transformation <French into Italy> for river network data

	Dataset FRANCE	**Transformation process**	**Dataset ITALY**
Dataset	**River network (HYLCOV00_rivers.shp)**		**River network (ELEMENTI_IDRICI.shp)**
Projection	Lambert-93 EPSG:2154	Coordinate transformation	GAUSS BOAGA - ROMA40 EPSG: 3003
Geometry	LINE		LINE
Attribute and Data Types	POSITION [string]	Rename attribute Classification Mapping *(e.g. French data POSITION=1 – Italian data SOTTOPASO=F)*	SOTTOPASO [string]

The harmonised data model (for two-side transformation) is based on specifications of INSPIRE as a key component of current SDI. Following specifications were mainly utilised:

- Hydrography data theme (INSPIRE Annex I) to exchange hydrological information (applied for dataset related to water network, e.g. watercourse, water bodies, etc.),
- Elevation data theme (INSPIRE Annex II) and Geographical Grid Systems (INSPIRE Annex I, 2) for Digital Terrain Model (altitude information necessary for watershed schematisation),
- Land cover data theme (INSPIRE Annex II) for land cover information influencing runoff,
- Environmental Monitoring Facilities data theme (INSPIRE Annex III) and Meteorological Geographical Features data theme (INSPIRE Annex III) for measurements (time series of water discharge, precipitation etc.).

INSPIRE based target schema can be seen on Fig. 8.

It is important to highlight that the data model is INSPIRE based but not fully INSPIRE adopted. This is done by the complexity of the INSPIRE and requirements to maintain more simple attribute implementation.

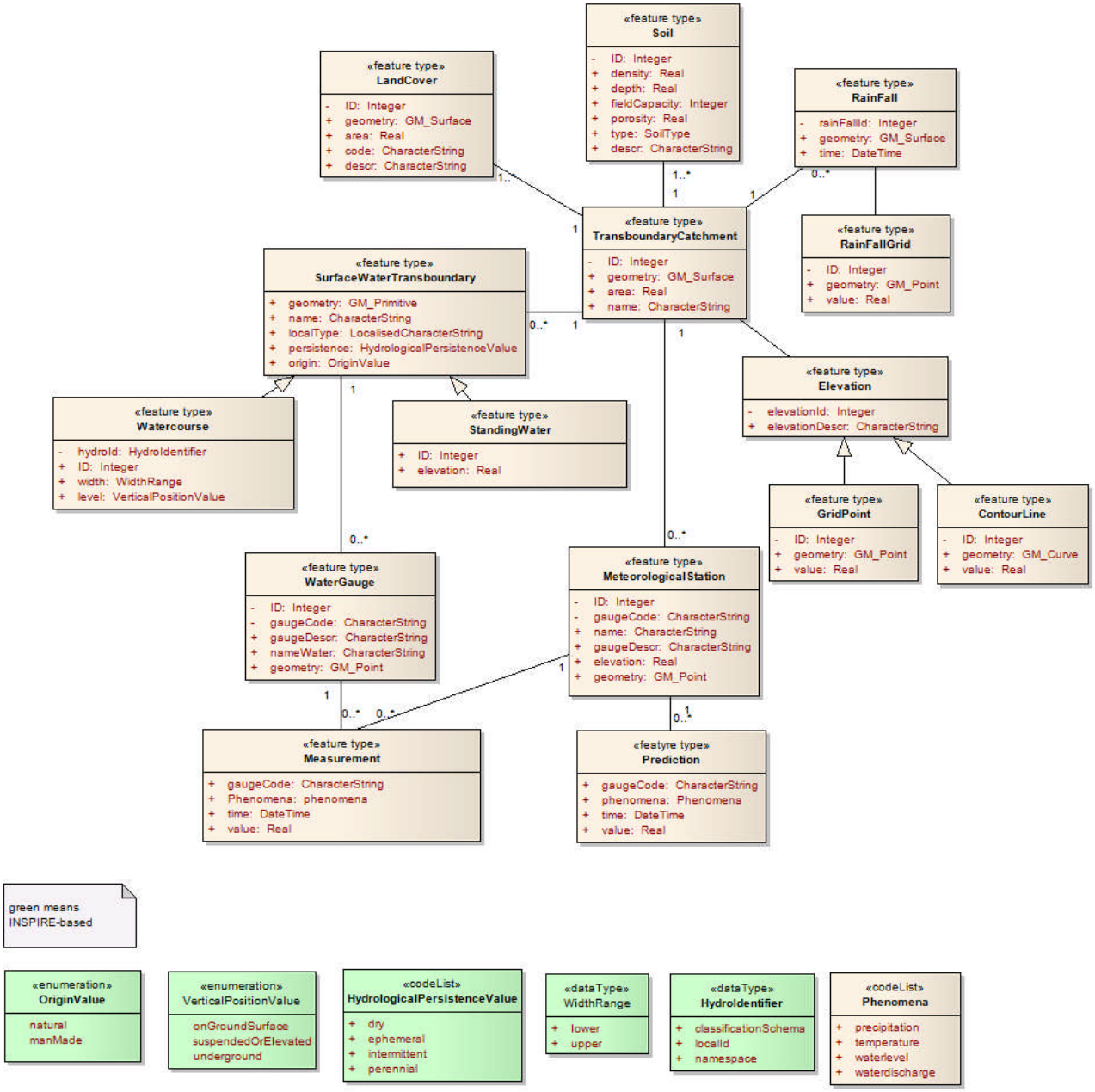

Fig 8: INSPIRE based target schema

HALE workflow

The following workflow description provides a list of required steps how to prepare transformation scheme using HALE. The description is based on the demonstrator prepared by GISIG (http://www.gisig.it/humboldt/training/level3/protected_areas/demonstrator/test1.html).

Loading Schemas and Data: The Source and Target schemas

The first step is to load the source and target schemas in the HALE Schema Explorer. We start with our "source" schema. In the current version of HALE, you can load any XML Schema, including metadata schemas, GML Application Schemas and others. However, the schema import is optimized for GML Application Schemas and supports the following versions: GML 2.1, GML 3.1 and GML 3.2. You can also load a schema from a shapefile (*.shp).

GML Application Schemas have to be available on a WFS server. It is possible to setup your own WFS server or use some existing one.

The source schema is imported from a Web Feature Service's GetCapabilities. To load the schema go to "File", "Open schema", select "Import as source Schema", press the “Change...” button and enter your server's Get- Capabilities request URL into the text field at the top of the appearing window. If the network requires the usage of one proxy server, it is necessary to configure the proxy server (use the Configure Menu, click “Network Settings”, enter your HTTP proxy host and port and click “Save settings”). After that all types will be loaded and shown in the list below the button. If used, WFS offers FeatureTypes from more than one namespace, it is now also required to pick one of the namespaces. Finally one sees the namespace of your schema in the left part of the Schema Explorer.

After loading our source schema it is possible to **load** also "**source data**" or “instance data”. This view enables to see a cartographic representation of the reference data for the source schema and the transformed data alongside each other, when you have loaded such data. It can be styled and navigated interactively. If the system cannot clearly identify the used CRS (Coordinate Reference System) from the data, it will request the user to provide either the EPSG code or the Well-Known Text for the used CRS.

Similarly the **target schema** is loaded. The target schema can be also derived from the description of the „home datasets“.

Now it is possible to **explore the source and target models** in the Explorer View. To have a good view of any large schema it is recommended to activate following options in the schema explorer:

- organize feature types as list to have a clear view of the features of the schema,
- repeat inherited properties to explore hidden parts,
- expand complex properties to have a clear vision of all elements.

Mapping schema items

The further step is mapping of the items. It means to build a mapping between source and target datasets (schema of classes/attributes changes). We start selecting the items (classes or attributes) we want to map in the Schema Explorer. Next we select a type of transformation during mapping. It is possible to check the details of proposed mapping in the “Mapping” window and split the map viewer to see the transformed geometry. The system offers to apply a specific style to the transformed data. It is recommended to use predefined matching table mappings and apply the transformations in the schema explorer selecting the appropriate mapping function. Usually attribute transformations are applied first.

Let us give an example of using HALE for harmonisation mapping for French Watercourse datasets.

Watercourse transboundary harmonisation

First we transform data from the French watercourse dataset called HYLCOV00_rivers to INSPIRE hydrography schema. After loading our source schema from Web Feature Service´s GetCapabilities (http://www.gisig.it:8081/geoserver/ows?service=WFS&request=GetCapabilities), we have to load target schema for transboundary catchment scenario in *.xsd format. Both schemas (source and target) can be seen in Fig.9. If you want to see your source data in Map Viewer, select File and Open Source Data from the toolbar.

The exploration of matching possibilities between the given dataset and INSPIRE hydrography schema revealed that four attributes have to be transformed. The remaining attributes of the source dataset are not needed and have to be excluded from the transformation process using INSPIRE Nil reason function.

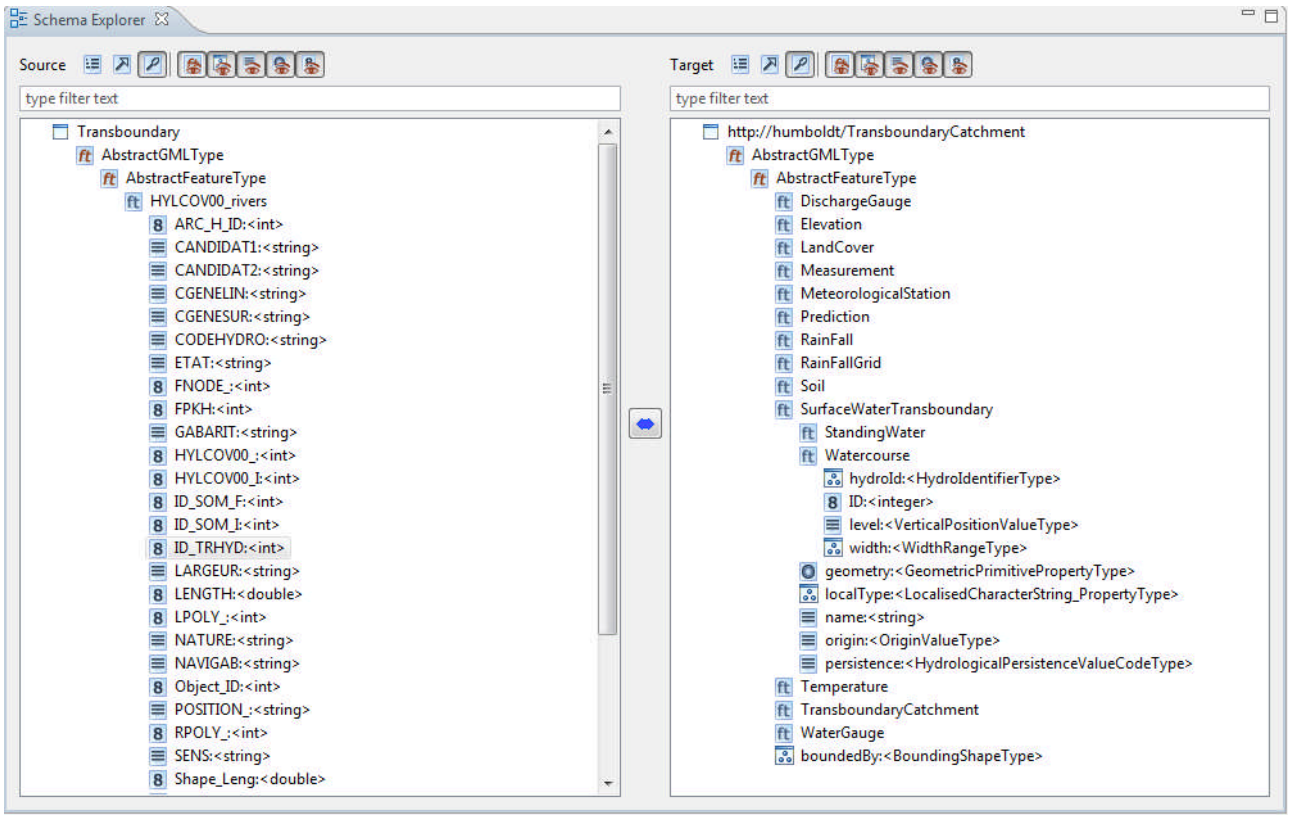

Fig 9: Source and Target schemas in HALE schema explorer

The attribute ID_TRHYD is equivalent to ID in INSPIRE schema (the target schema) (see Fig 9). We use „Rename function“ for this transformation. The transformation is repeated for attributes LARGEUR (rename to WIDTH), NATURE (rename to ORIGIN) and POSITION (rename to LEVEL).

It is possible to review the results of your transformations within HALE in the Mapping window (Fig.10).

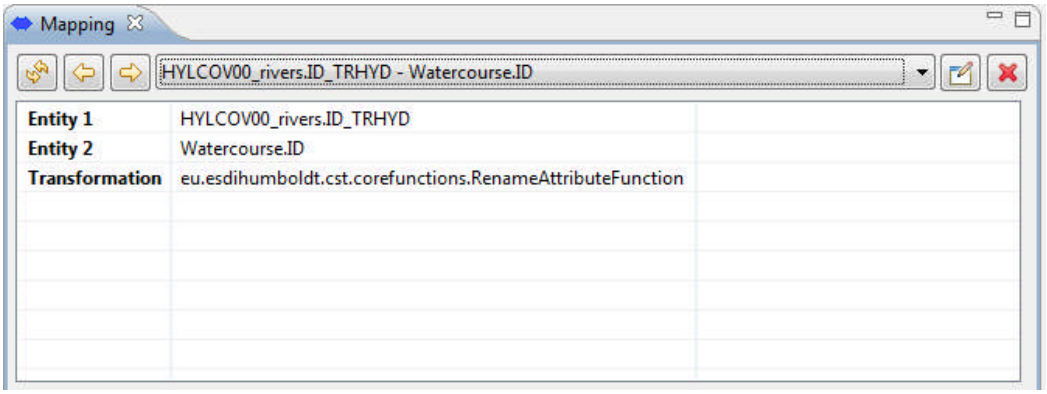

Fig10 Result of renaming function in HALE Mapping window

Next step is to classify some values inside attributes that are mandatory in the watercourse attribute subset of the INSPIRE schema. Classification mapping function allows to map values of a source attribute to fixed values of a target attribute (to reclassify values of a source attribute to the required values of the target). The relation is always a many to one relation, and each code from the source schema can only be mapped once. The function “Classification mapping” is applied to replace values in the attributes LARGEUR, NATURE and POSITION. Matching table can be seen in Table 6.

Table 6 Matching table

France	INSPIRE
LARGEUR 1: from 0 to 15m 2: from 15 to 50m 3: more than 50m	Width width=lower width=upper width=upper
NATURE NATURE=1 NATURE=3 or 4	Origin origin=natural origin=manmade
POSITION POSITION=2 POSITION=1 POSITION=3	Level level=SuspendedorElevated level=onGroundSurface level= underground

Now select both attributes, click the central arrow, and run the “Classification Mapping” function. Select value of target schema from the list and add old value from the source schema in Classification Mapping window. Classification Mapping window is shown in Fig.11. Repeat the procedure for all attributes which you need to classify. The results of classification mapping function can be revised in HALE Mapping window.

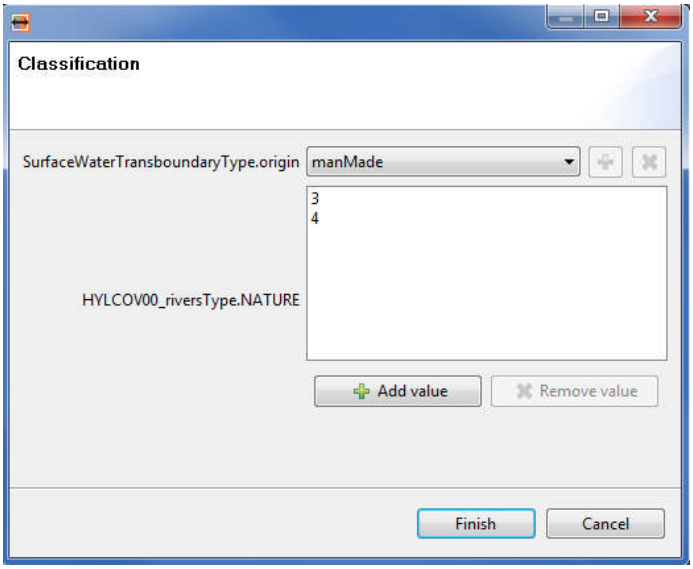

Fig 11: Classification mapping window

When we have no data for some attributes it is recommended to use „Attribute Default Value" function to fill the mandatory fields and „INSPIRE Nil reason“ function for optional attributes. The first function fill the whole attribute with a defined value. The later function sets the attribute unpopulated.

Saving the alignment project

After finishing all mappings it is necessary to save the alignment project. This saves an XML and a GOML file with the same name in the same directory or to an alternative mapping file and to an alternative place. GOML is required to make a corresponding data transformation.

Schema translation and transformation to the target schemas

The harmonised schema (described by HALE and saved in an OML file) is used by CTS to implement the required process. The Conceptual Schema Translation Service (CST) is a Web Processing Service for transforming data from one application schema to another (CST, 2010). Note that this tool is currently in testing phase. CST is Java library that is responsible for

- parsing and generating GML,
- parsing OML and
- execution of particular transformation of spatial features.

CST also provides WPS interface for executing the transformation. For this propose pyWPS library was reused where a Java – python binding represents a new contribution. CST internally uses GeoTools library for representing the feature model. CST -WPS provides OGC complaint WPS interface. For simple access to this interface there is also HTML Client (Fig. 12).

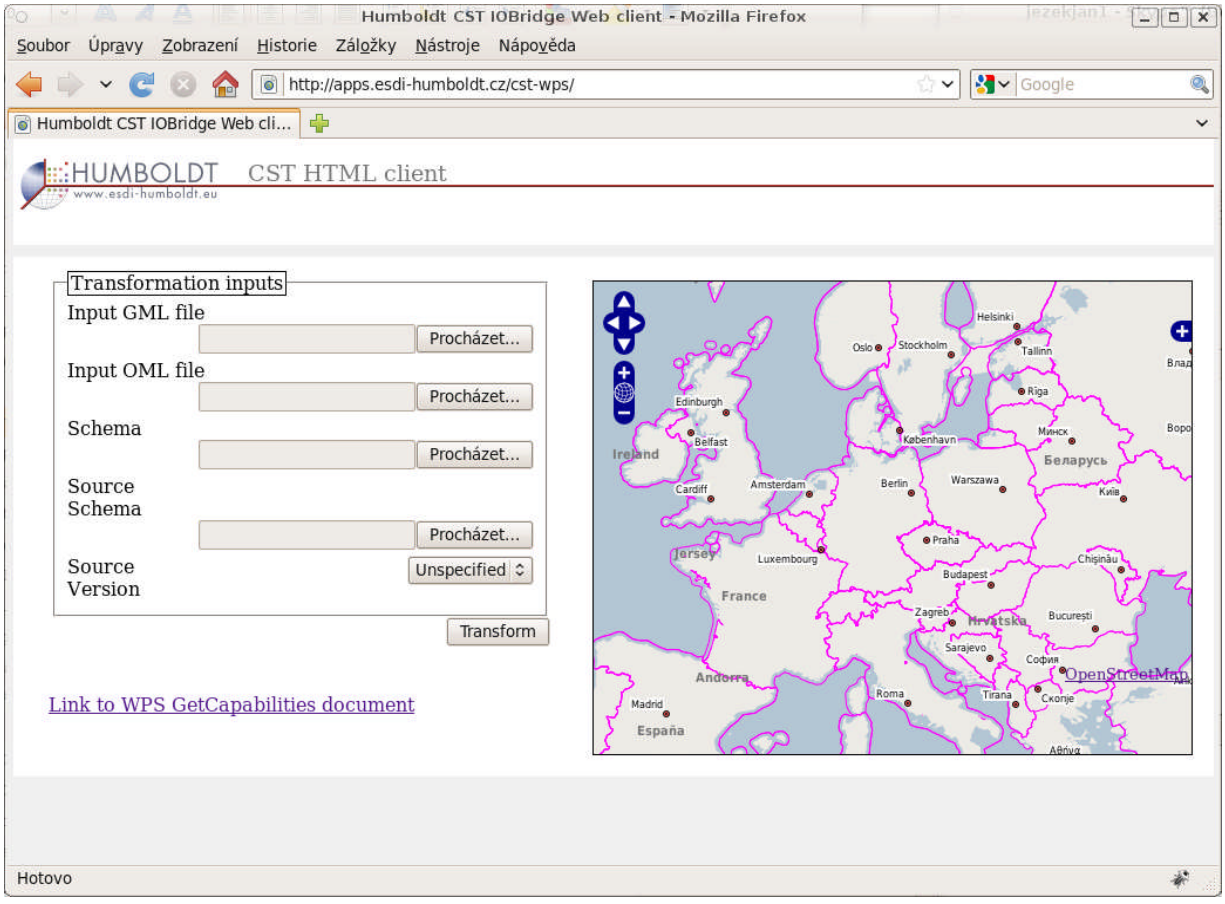

Fig 12: HTML client for WPS interface to CST

DISCUSSION AND CONCLUSION

The HUMBOLDT project contributes to the developments in data harmonisation using web tools and local designing tools, enabling appropriate transformation to create harmonised data sets. As discussed in the paper, such tools are critical for the current implementation phase of INSPIRE. Our experiments have shown that HALE and CST are very promising developments, which close a large gap in the market, i.e. conceptual schema mapping and automatic data transformation. As shows in our paper it is possible to design either one-side transformation (datasets from foreign country to match local datasets) or two-side transformation (where data from both countries are transformed into a common target schema) using HALE. The

transboundary scenario is a good demonstration since many of the mentioned data sets are relevant for other cases and scenarios.

Nevertheless, some issues still obstruct full implementation of all proposed tools. One of these drawbacks in the INSPIRE implementation process is the requirement of GML 3.2.1 which is not widely adapted in current applications. Even more existing implementations of GML are not consistent. It may be a critical issue for the wide utilisation of the Humboldt framework.

The Humboldt project is not alone in the aim of facilitating data harmonisation. Standardization of relevant data structures is a subject of many projects. Another important European initiative is the WISE project (http://water.europa.eu/), which provides a repository for a wide range of GIS datasets. These datasets can be compiled by Member States for regulatory reporting as well as the WISE Reference GIS datasets. Information about the reference GIS datasets and data models for themes connected with EU water related directives can be found in documentation of this project namely guidance document No. 22. These data sets can also be used as target schemas when harmonising data sets.

Humboldt framework (based on web services) address successfully future system requirements, especially web portals. The group of water related web services is still growing. TRANSCAT Decision Support System T-DSS (Horak et al., 2006) was one of the first modular web application system build on open sources technologies aimed at water management and utilisation of hydrological modelling. The European project HarmonIT (Gijsbers et al., 2004) addressed issues of spatial and temporal scale differences, unit differences etc. They launched an Open Modelling Interface (OpenMI, http://www.openmi.org) enabling seamless interaction among modelling systems, the integration and combination of their functions. CUAHSI (the Consortium of Universities for the Advancement of Hydrologic Science) community provides a group of web services called WaterOneFlow. CUAHSI web services facilitate the retrieval of hydrologic observations information from online data sources using the SOAP protocol. CUAHSI WaterML is an XML schema defining the format of messages returned by the WaterOneFlow web services (Zaslavsky et al, 2007).

REFERENCES

Bedient, P. and Huber, W. (2001) Hydrology and floodplain analysis, 2nd edition, Prentice Hall, London, 763 pp.

Beven, K. (2002) Rainfall-runoff modelling, The Primer, John Wiley, London, 372 pp.

Blasby D., Davis M., Kim D., Ramsey P. (2004) GIS Conflation using Open Source Tools. A white paper of JUMP project. pp. 32. http://www.jump-project.org/assets/JUMP_Conflation_Whitepaper.pdf

Charvát K., Kocáb M., Konečný M., Kubíček P. (2007): Geografická data v informační společnosti. VÚGTK 2007, 269 pp. ISBN 978-80-85881-28-8.

CST (2010), Conceptual Schema transformer. http://agile2010.dsi.uminho.pt/pen/ShortPapers_PDF/112_DOC.pdf <On-line> Cited 30.10.2010.

EUREF (2010), the IAG Reference Frame Sub-Commission for Europe: http://www.euref.eu/. On-line. Cited 20.10.2010.

Fonseca F.T., Egenhofer M.J., Davis Jr. C. A., Borges K. A. V. (2000) Ontologies and Knowledge Sharing in Urban GIS, Computer, Environment and Urban Systems, vol. 24, pp. 251-271

Galdos Systems Inc.: GML Complexity. Pp.4, <On-line> http://www.galdosinc.com/archives/186, cit. 28.10.2010

Guidance Document No. 22 - Updated Guidance on Implementing the Geographical Information System (GIS) Elements of the EU Water legislation: http://circa.europa.eu/Public/irc/env/wfd/library?l=/framework_directive/guidance_documents&vm=detailed&sb=Title. Appendices to the guidance are available on EEA CIRCA at: http://eea.eionet.europa.eu/Public/irc/eionet-

circle/eionettelematics/library?l=/technical_developments/wise_technical_group/updated_2ndedition/appendices_updated&vm=detailed&sb=Title

Horák J., Orlík A., Stromský J. (2008) Web services for distributed and interoperable hydro-information systems. Hydrol. Earth Syst. Sci., 12, 635-644. ISSN 1027-5606. http://www.hydrol-earth-syst-sci.net/12/635/2008/hess-12-635-2008.pdf

Horak J., Unucka J., Stromsky J., Marsik V., Orlik A (2006) TRANSCAT DSS architecture and modelling services. Control & Cybernetics, vol 35, No.1, Warsaw 2006, ISSN 0324-8569, 47-71.

Institut Géographique National (Ed.) (2002) BD CARTHAGE Version 3.0 - Descriptif de contenu (sphère eau), edition 1 (2002).

Lehto, L., 2007, Schema Translations in a Web Service Based SDI, 10th AGILE International Conference on Geographic Information Science 2007Aalborg University, Denmark,

Lehto, L. and T. Sarjakoski, 2004. Schema translations by XSLT for GMLencoded geospatial data in heterogeneous Webservice environment. Proceedings of the XXth ISPRS Congress, July 2004, Istanbul, Turkey, International Archives of thePhotogrammetry, Remote Sensing and Spatial Information Sciences, XXXV (B4/IV):177182, also CDROM.

Lutz M, and E. Klien (2006) Ontology-based retrieval of geographic information. International Journal of Geographical Information Science. Vol. 20, No. 3, March 2006, 233-260.

Maidment, D. (1993) Handbook of Hydrology, 1st edition, McGraw Hill Professional, London, 1424 pp.

Mitiku, A. B., (2009) Modelling of International Transboundary Catchment Area - Roya Basin and Its Data Harmonization Needs, Master thesis, University of Nice Sophia Antipolis. 104 p.

Národní geoinformační infrastruktura České republiky (2001) Program rozvoje v letech 2001-2005. NEMOFORUM, Praha. [on-line] http://www.cuzk.cz/GenerujSoubor.ashx?NAZEV=999-NGII

Nougeras-Iso, J., Medrano, P., Zarazaga-Soria, P. (2005) Geographic Information Metadata for Spatial Data Infrastructures, Springer, Boston.

Pauknerová, Tryhubová (2006) INSPIRE and geoinformation infrastructure in the CR. In Internet ve státní správě a samosprávě. Hradec Králové, 3-4.6.2006, VUGTK.

Reitz, T. (2010a) HUMBOLDT Alignment Editor Manual 2.0.M2. <On-line>

http://community.esdi-humboldt.eu/attachments/72/alignment_editor_manual_2010-03-31-M2.pdf

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

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

Sheth A (1999) Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In M.F. Goodchild, M. Egenhofer, R. Fegeas and C.A. Kottman (Eds), Interoperating Geographic Information Systems, pp. 530 (Dordrecht, Netherlands: Kluwer Academic).

Stuckenschmidt H. (2003) Ontology-Based Information Sharing in Weakly Structured Envirnoments. PhD thesis, Vrije Universiteit Amsterdam.

Šeliga M., Růžička J. (2005) Geoinformation systems and web services. Acta Montanistica Slovaca. 2/2005, Technical University of Kosice, the Faculty of Mining, Ecology, Process Control and Geotechnologies (F BERG), Košice 2005, ISSN 13351788. p. 206-212

Vaccari L., Shvaiko P., Marchese M. (2009) A geo-service semantic integration in Spatial Data Infrastructures. International Journal of Spatial Data Infrastructures Research, 2009, vol.4, 24-51.ISSN 1725-0463.

Vries, de, M., Giger, Ch., Iosifescu, I., Laurent, D., Althoff, J.S. (2010) A7.0 D1 Concept of Data Harmonisation Processes. HUMBOLDT project deliverable.

Zaslavsky, I., Valentine, D., Whiteaker, T. (Ed.) (2007) CUAHSI WaterML, Open Geospatial Consortium Inc. http://portal.opengeospatial.org/files/?artifact_id=21743