International Journal of Spatial Data Infrastructures Research, 2011, Vol.6, 234-260

Data Harmonisation Put into Practice by the HUMBOLDT Project*

Astrid Fichtinger[1], Joachim Rix[3], Ulrich Schäffler[1],
Ines Michi[2], Moses Gone[3] and Thorsten Reitz[3]

[1]Technische Universität München (astrid.fichtinger@bv.tum.de;
ulrich.schaeffler@bv.tum.de)

[2]Technische Universität Darmstadt (ines.michi@gris.informatik.tu-darmstadt.de)

[3]Fraunhofer Institute for Computer Graphics Research
(joachim.rix@igd.fraunhofer.de; moses.gone@igd.fraunhofer.de;
thorsten.reitz@igd.fraunhofer.de)

Abstract

Data harmonisation is a key prerequisite for an efficient and meaningful combination of heterogeneous information in cross-border applications and spatial data infrastructures. This is also the main objective of the INSPIRE Directive which has entered its implementation phase. Data Specifications for INSPIRE Annex I data themes have been published containing harmonised, pan-European data models and a number of other requirements. Data providers across Europe face the challenge of transforming their legacy data to comply with these Data Specifications. This paper presents results of the European project HUMBOLDT. Data harmonisation requirements identified in nine scenarios covering a wide range of application domains and using heterogeneous data from a number of European countries are illustrated. Processes required to achieve data harmonisation are described from an application point of view. The open-source software framework for data harmonisation and services integration developed in the project is introduced and its use in two application scenarios is demonstrated.

Keywords: HUMBOLDT, INSPIRE, data harmonisation, schema translation

---

DOI: 10.2902/1725-0463.2011.06.art11

1. INTRODUCTION

"To satisfy increased demands for the use, sharing, and exchange of geographic data in cross border European applications, methods to support interoperability are required by the community" (Friis-Christensen et al, 2005).

The INSPIRE Directive has added momentum to this trend in the geospatial community. Now, the Directive has entered implementation phase and Data Specifications for INSPIRE Annex I data themes have been published. These contain harmonised, pan-European data models and a number of other requirements, e.g. encoding rules and spatial reference systems to use. Data providers across Europe face the challenge of transforming their legacy data to comply with these Data Specifications. This is a complex process involving a variety of data harmonisation issues, which today hamper a cross-domain, cross-country and pan-European exchange of spatial information. Efficient, cost-effective and user-friendly tools for different steps in the data harmonisation process are needed. The approaches, processes and tools have to be tested involving data currently in use at mapping or environmental agencies across Europe.

The European project HUMBOLDT[1] (www.esdi-humboldt.eu) running from 2006 to 2011 focuses on these aspects of data harmonisation with a view to the implementation of INSPIRE, but open for other application fields as well. 27 partners from 14 European countries representing public administration, research and industry have teamed up to contribute to the implementation of the European Spatial Data Infrastructure (ESDI). The project commenced with a comprehensive state-of-the-art analysis on methods and tools for spatial data and metadata management and harmonisation (cf. de Vries et al, 2007a). Suitable software architectures as well as processes were described and user requirements of a variety of application domains were gathered. This analysis serves as a basis for the continuous development of a software framework ("HUMBOLDT Framework") including tools and services to support spatial data and service providers in offering standardised spatial information by creating partly automated, but individually adjustable processes for the harmonisation of spatial data and metadata. Challenging data harmonisation steps, for which no ready-made solution exists, such as conceptual schema translation, were subject to thorough investigation and led to the development of new software solutions. The solutions developed within the project are available as Open Source Software at http://community.esdi-humboldt.eu/ for access by and support of the community.

---

[1] The HUMBOLDT project is funded by the EU under the 6th Framework Programme - Aeronautics and Space (GMES) Thematic Priority.

An important driver for the whole software development process in HUMBOLDT is the establishment of application scenarios from which requirements were drawn and which demonstrate the applicability of HUMBOLDT Framework components. Nine scenarios were set up based on real-world use cases covering a wide range of INSPIRE and GMES[2] related application domains (air quality, border security, flood risk management, forest and urban planning, oil spill monitoring, protected areas, sustainable urban atlas, transboundary catchments[3]). Heterogeneous data from a number of European countries (e.g. Austria, Czech Republic, France, Germany, Greece, Italy, Hungary, Portugal, Spain, Switzerland and the United Kingdom) is used in the scenarios. Since these scenarios are linked to real-world applications, their results from using and evaluating the Framework and its toolset provide essential information on its user-friendliness, on how it addresses users' needs and matches the requirements for further development. Moreover, the gathered information is valuable for guidelines and best practice examples on how tools and standards can be used to create the ESDI. Therefore it has also been documented in training material which can be accessed via a web-based training plaform at http://www.gisig.it/humboldt/training/.

The paper gives an overview of data harmonisation in the HUMBOLDT project from an application point of view. It focuses on the technical aspects of data harmonisation (e.g. schema translation) rather than legal or organisational aspects. First, data harmonisation requirements and user needs are addressed, followed by a description of the harmonisation process. In the next steps the HUMBOLDT Framework with its Tools and Services is introduced and their use for two different scenarios is explained. One of the scenarios, named European Risk Atlas (ERiskA), aims at developing a cross-border flood risk management application for the Lake Constance region which includes Swiss, Austrian and German territories. The second scenario, Atmosphere, demonstrates possibilities to provide users with air quality information adapted to their needs within a mobile environment. The final section will conclude the results and benefits of the harmonisation efforts and solutions, as well as provide an outlook on future perspectives of these developments.

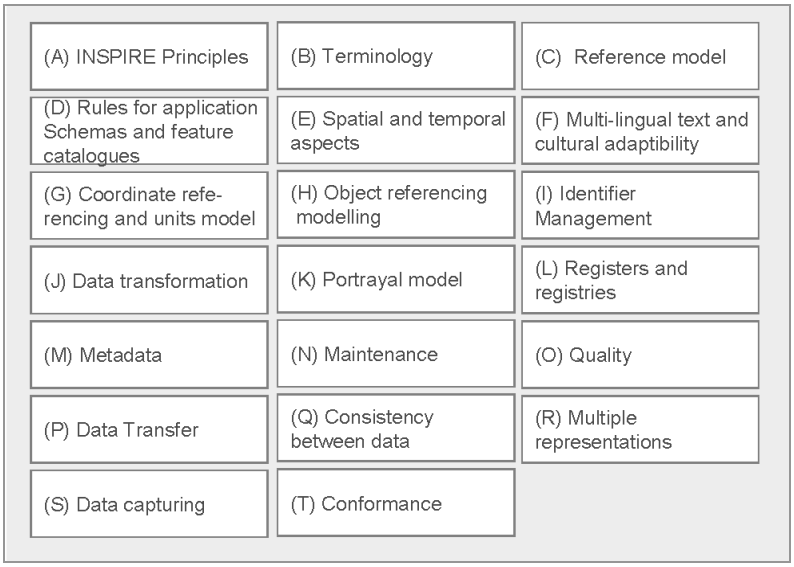
2. DATA HARMONISATION REQUIREMENTS

As data harmonisation comprises many different aspects, there are also multiple ways to define concepts related to data harmonisation, depending on which aspects are in the focus.

[2] Global Monitoring from environment and Security, http://www.gmes.info/
[3] please refer to http://www.esdi-humboldt.eu/scenarios.html

In the INSPIRE Directive, the term “interoperability” is used, meaning “the possibility for spatial data sets to be combined, and for services to interact, without repetitive manual intervention, in such a way that the result is coherent and the added value of the data sets and services is enhanced” (Commission of the European Communities, 2007). INSPIRE identifies 20 different aspects relevant for data harmonisation (“data interoperability components”, see figure 1) which have to be covered by the provisions in the INSPIRE implementing rules and technical guidelines.

Figure 1: INSPIRE Data Interoperability Components



Source: Drafting Team Data Specifications, 2008

This comprehensive list deals with many different aspects of data harmonisation and interoperability in a spatial data infrastructure. There are data model related issues like rules for application schemas or spatial and temporal aspects as well as issues related to the data instances themselves like spatial reference systems, data quality and consistency, e.g. at borders or in cases where there are multiple representations of the same real-world object. In addition to that, aspects related to data capturing and maintenance as well as visualisation (portrayal) are covered.

Not all of the above-mentioned components are within the scope of the HUMBOLDT project in which data harmonisation is defined as “creating the possibility to combine data from heterogeneous sources into integrated,

consistent and unambiguous information products, in a way that is of no concern to the end-user" (de Vries et al, 2007a).

The following data harmonisation issues have been identified in the different HUMBOLDT application scenarios (de Vries et al, 2007a, 2007b):

- Data formats (e.g. raster/vector; proprietary formats of different vendors like ESRI shapefiles)
- Spatial reference systems (e.g. different projections, coordinate systems, datums and ellipsoids)
- Conceptual schemas/data models (e.g. different modelling methods, modelling languages and structures)
- Classification schemes (e.g. different ways to classify land cover or flood risk warning levels)
- Scales/resolutions of vector or raster datasets respectively (e.g. resolutions of digital terrain models varying from 1 to 50 metres)
- Levels-of-detail of data content (e.g. only roads or also walkways)
- Metadata (e.g. different metadata profiles or lack of formalised/standardised metadata)
- Terminology (e.g. different application domains and natural languages)
- Portrayal (e.g. different styles in map symbology)
- Multiple representation of the 'same' spatial objects (e.g. when data sets overlap or different levels of generalisation are present)
- Spatial consistency at borders (e.g. edge-matching)
- Processing functions (parameters and computational functions/algorithms, e.g. to derive forecast models, or to compute land cover classification based on remotely-sensed images)

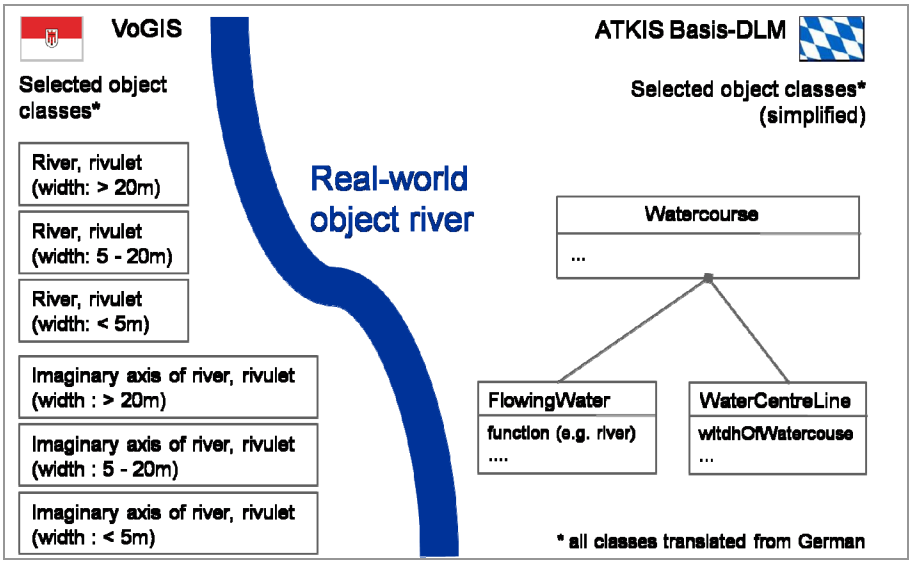
Solutions to overcome heterogeneities related to conceptual schemas are among the ones with the highest priority in the project. Those heterogeneities can be categorised e.g. in the following way (Stuckenschmidt, 2003 as cited in Friis-Christensen et al, 2005):

1. Syntax (related to different data formats),
2. Structure (related to differences in conceptual schemas),
3. Semantics (related to differences in the intended meaning of terms).

The syntax heterogeneities can be dealt with today – to a certain degree – using standardised web services like the Open Geospatial Consortium's Web Feature Service (WFS) which encapsulates the internal structure of the data. Interoperability issues can still arise from different versions of the standard or vendor specific implementations. In the HUMBOLDT project, the focus lies on the structural and partly also the semantic heterogeneities. In many cases, there is a lack of conceptual data models describing data in a formalized way using a

conceptual schema language in the HUMBOLDT application scenarios. It is also quite common that similar real-world objects such as transport networks or hydrographic features are modelled very differently in different countries or application domains. In some cases this also involves different conceptual schema languages or different profiles of the Unified Modelling Language (UML), e.g. the international ISO or the Swiss INTERLIS profile. Figure 2 illustrates the different ways used to structure the real-world object river in topographic vector data provided by the mapping agencies of the Austrian state of Vorarlberg (left hand side) and the German state of Bavaria (right hand side).

Figure 2: Different Ways to Structure the Real-world Object River



Source: based on Fichtinger and Kutzner (2010)

The specifications for INSPIRE Annex I data themes contain harmonised UML conceptual schemas and GML application schemas as well as feature catalogues. Data providers across Europe have to provide their data compliant with the data specifications. Since it is not required by INSPIRE – and probably also not reasonable in many cases – that the data providers change the ways their legacy data is modelled, structured and stored, mappings are required between the legacy schemas and the INSPIRE schemas. This of course also applies to other harmonisation tasks beyond INSPIRE in spatial data infrastructures with other target schemas.

The functions required for these mappings can be grouped e.g. according to a classification by Lehto (2007) (examples of mapping of data from Vorarlberg to

INSPIRE schemas taken from the ERiskA scenario are used for illustration purposes):

1. Filtering of e.g. features based on values of an attribute (see figure 3) or of attributes, meaning that based on a conditional statement only selected features or attributes are mapped to the target schema.
2. Renaming of features, attributes and their values (e.g. translation between languages)
3. Reclassification of features or attribute values (e.g. converging to a coarser classification system of land use in the target attribute value)
4. Merging/splitting of features (e.g. merging of watercourse segment features to form one watercourse feature) or attributes (e.g. concatenation of two or more attribute values in the source to form a single attribute value in the target schema)
5. Value conversions, either geometric (e.g. polygon to line) or alphanumeric (e.g. units of measurement conversion from miles to kilometres)
6. Augmentation (e.g. deriving values for target schema attributes which are missing in the source schema based on values of other attributes in the source schema or filling in default values, see figure 4).

Figure 3: Example of Filtering Features based on Attribute Values

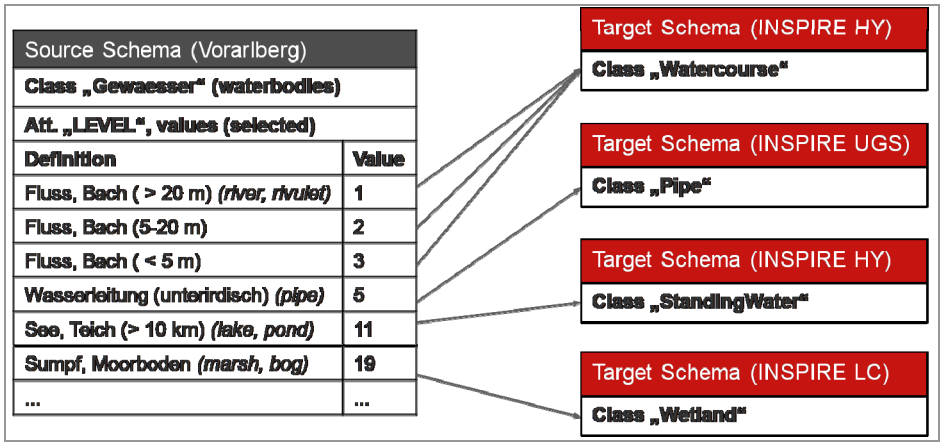
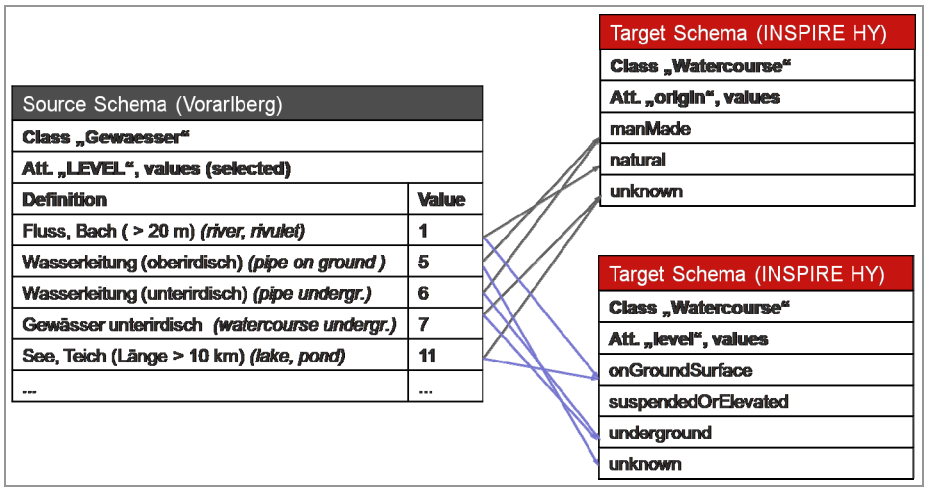


Figure 4: Example of Augmentation Features based on Attribute Values



Further harmonisation issues in the cross-border HUMBOLDT scenarios include e.g. spatial reference systems and spatial consistency at borders. Figure 5 shows the situation in the Lake Constance region using topographic vector data on watercourses from national/state mapping agencies in the region. Within this relatively small geographic area, two different scales and four different spatial reference systems are used:

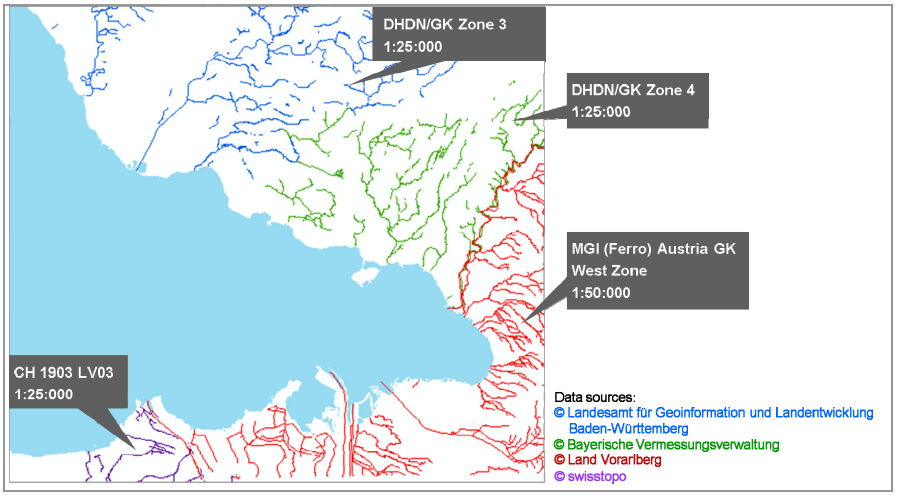
- in the German state of Baden-Wuerttemberg (blue): Deutsches Hauptdreiecksnetz / Gauss-Krueger Zone 3" (DHDN/GK Zone 3), EPSG code 31467
- in the German state of Bavaria (green): "Deutsches Hauptdreiecksnetz / Gauss-Krueger Zone 4" (DHDN/GK Zone 3), EPSG code 31468
- in the Austrian state of Vorarlberg (red): Militar-Geographisches Institut (Ferro) / Austria Gauss-Krueger West Zone (MGI (Ferro) Austria GK West Zone), EPSG code 31251
- in Switzerland (purple): CH1903 / Landesvermessung 1903 (CH1903 / LV03), EPSG code 21781

Figure 6 (a detail of figure 5) shows several spatial inconsistencies at the border of Bavaria and Vorarlberg:

- Multiple representation of the border river Leiblach with different spatial representation: polygon feature in the Bavarian dataset and line feature in the dataset from Vorarlberg

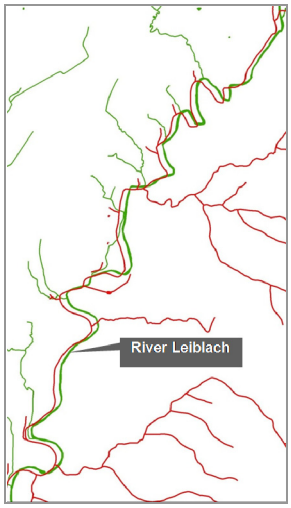
- A systematic shift of position caused by re-projection using standard parameters of a desktop GIS software as well as small differences in position presumably due to differences in data capture and generalisation.

Figure 5: Different Spatial Reference Systems and Scales in the Lake Constance Region



Source: based on Fichtinger and Kutzner (2010)

Figure 6: Spatial Consistency Issues at the Border (detail from figure 5)



The selection of data harmonisation requirements presented here is based on the analysis of the different HUMBOLDT application scenarios and is thus not exhaustive. Nevertheless, it can be assumed that these are typical issues to be dealt with also in other cross-border application scenarios. The following chapters describe potential solutions meet these requirements.

3. DATA HARMONISATION SOLUTIONS

3.1. Data Harmonisation Process

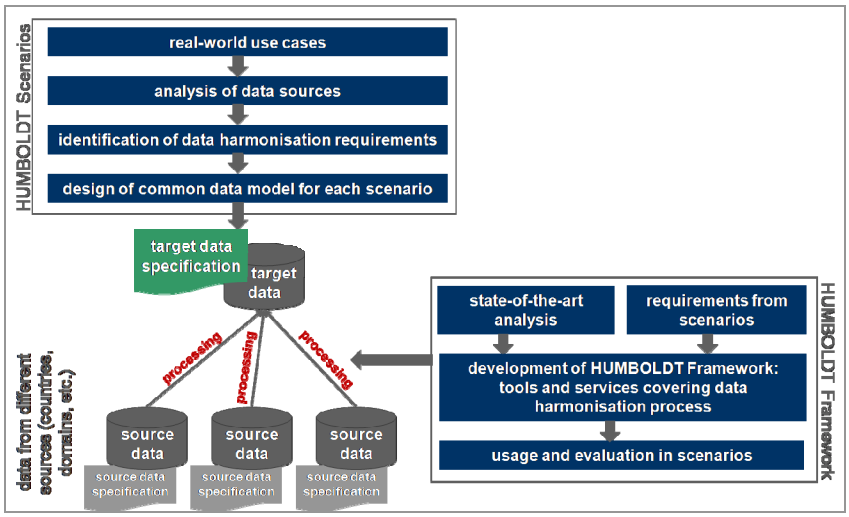
Two main phases can be distinguished within a data harmonisation process:

1. Definition of the target (e.g. a specification comprising a conceptual schema and other data characteristics)
2. Processing (e.g. transforming heterogeneous data from different sources to match the target data specification)

The RISE (Reference Information Specifications for Europe) project has proposed a methodology and guidelines (RISE, 2007) for the first phase which has also served as a base for creating the INSPIRE methodology for the development of data specifications (Drafting Team "Data Specifications", 2008). Both documents identify the following steps to be undertaken when creating a data specification:

- Development and description of use cases
- Identification of user requirements and spatial object types for use cases
- As-is analysis (analysis of existing data) including data harmonisation issues
- Gap analysis
- Development of data product specification including application schema
- Implementation, test and validation.

Figure 7: The Data Harmonisation Process



This approach has been tested in the HUMBOLDT project (see figure 7). Although no formal data specifications according to ISO 19131 have been created in the project, the methodology has proven useful, since it includes general aspects that can be transferred to the development of the HUMBOLDT scenarios. In a first step, real-world use cases for different application domains (see chapter 1) with relevance for INSPIRE and/or GMES have been developed. They were described using a template based on the RISE use case template (RISE, 2007) and UML use case diagrams. Domain experts and developers cooperated to identify requirements and describe the use cases in a specification document for each of the HUMBOLDT scenarios. Subsequently, a thorough information analysis was undertaken for each scenario. Information items needed for the use cases were identified and available datasets were analysed. This was documented in a “data profile” document for each Scenario as well as a table describing in detail the characteristics of the datasets/web services (e.g. formats, attributes, geometry types, metadata profiles, spatial reference systems, scales/resolutions, languages, etc.) used in the scenario. By comparing the results of this “as-is analysis” to the requirements, data harmonisation issues were described, also taking into consideration the checklist for data interoperability provided by RISE/INSPIRE (Drafting Team "Data Specifications" 2008). Based on the previous steps, a common conceptual data model was designed for each scenario containing agreement on classes, attributes, associations, code lists and enumerations, constraints and methods to be used. This was done in an iterative process, taking into consideration the feedback of stakeholders and developers. The concepts, requirements and recommendations for this step are described in de Vries (2007b). The data models were formalised

in UML application schemas using different UML editors (existing commercial and open source tools as well as the “GeoModel Editor” developed in the HUMBOLDT project). GML application schemas were derived from the UML application schemas according to ISO 19136.

Where possible, existing application schemas for INSPIRE spatial data themes - e.g. Administrative Units, Geographical Names, Hydrography, Natural Risk Zones (draft), Protected Sites and Transport Networks were re-used. Figure 8 contains a small clipping from the UML data model designed for the HUMBOLDT ERiskA Scenario. In this case, several feature types of the INSPIRE application schemas “Hydrography – Physical Waters” and “Natural Risk Zones” were re-used. This was done following the guidelines for extensions to INSPIRE application schemas given in the INSPIRE Generic Conceptual Model document (Drafting Team "Data Specifications", 2009). Since the existing INSPIRE application schemas are not to be changed, an additional, separate application schema containing several packages (e.g. ERiskA_HY for water-related features) is created for ERiskA. The ERiskA_HY package imports the above-mentioned INSPIRE application schemas (see figure 9). In case additional attributes are needed for ERiskA feature types, extended feature types (e.g. “Watercourse”) are created within the new ERiskA_HY application schema as subtype of the respective INSPIRE feature type. In addition to that, new ERiskA feature types (e.g. Gauge) are created as subtypes of the abstract INSPIRE HydroObject feature type.

In the second phase of the data harmonisation process different processing steps have to be executed to transform heterogeneous data from different sources to match the target data specification. Processing steps may include but are not limited to transformation of data from source to target conceptual schema, coordinate transformation, edge matching, language transformation, etc. Tools and services for different data harmonisation tasks developed within the HUMBOLDT project are described in section 3.2. Section 3.3 describes how these tools and services are applied in two of the HUMBOLDT Scenarios.

A prerequisite for conceptual schema transformation is the definition and formalisation of mappings - i.e. rules for transformations between source and target schemas. This can be done on different levels of data modelling (e.g. conceptual schema or data format level) using different mapping languages (proprietary or open/standardised) and tools assisting the user in creating the mapping rules. A comprehensive overview of existing approaches and projects is given in de Vries et al (2007a). In many projects, scripts/software have been written for a specific harmonisation process, the rules being hard-coded in the software code. In other projects, mapping rules are formalised separately from the transformation software code using a mapping language and thus allowing for

more flexibility and sustainability, e.g. in case the source or target schemas changes.

Figure 8: ERiskA Common Data Model (detail) with INSPIRE Feature Types (beige) and ERiskA Extension (blue)

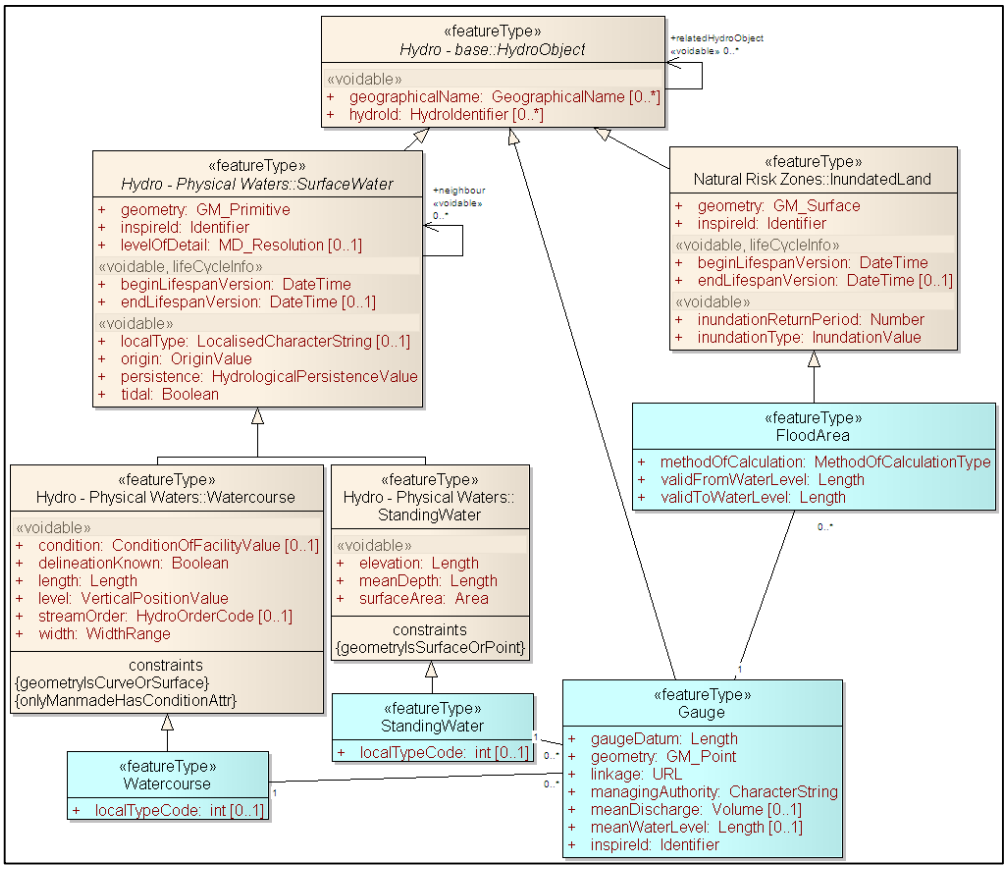
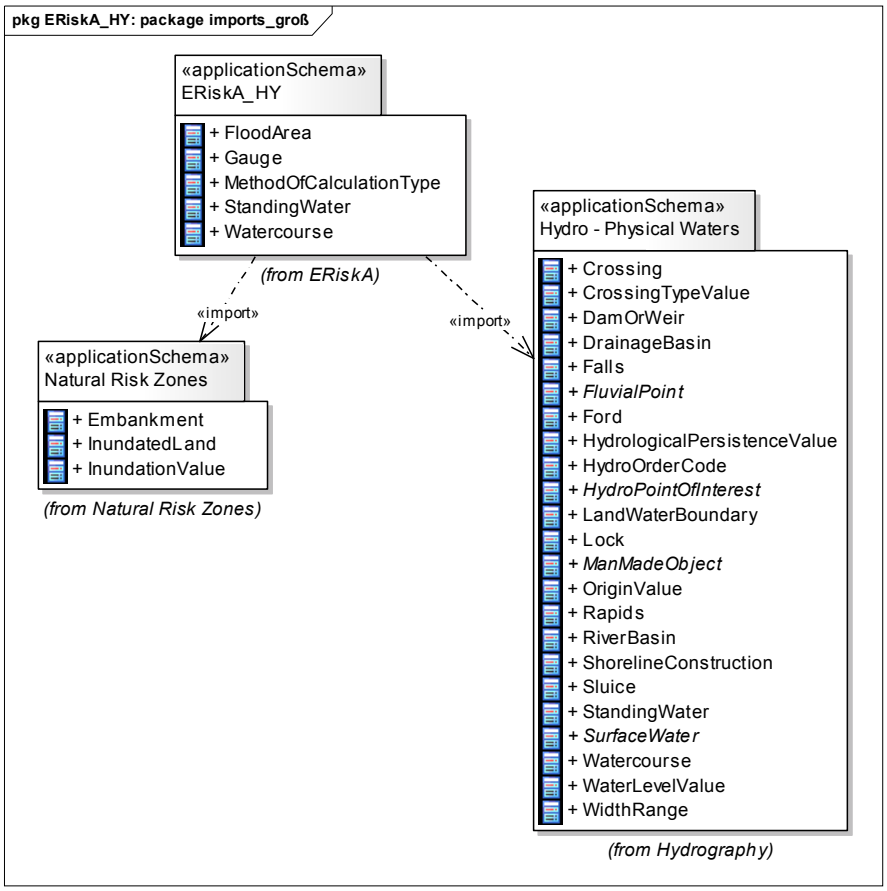


Figure 9: ERiskA Common Data Model (Package Imports)



Examples are XSLT (Extensible Stylesheet Language Transformation) adapted for mappings between GML application schemas used e.g. in the GiMoDig project (Lehto, 2007) and also tested in the HUMBOLDT project, INTERLIS used in Switzerland or UMLT used in the mdWFS project for mappings on the conceptual level between UML data models (Donaubauer et al, 2007). However, there is currently no commonly agreed standard for a mapping language in the geospatial community. In the HUMBOLDT project the following requirements amongst others were identified for such a mapping language:

- Description of mapping rules in a formal, unambiguous and machine-readable way
- Open (non-proprietary)
- Generic (implementation-neutral)
- Expressive enough to cover all transformation functions needed

After evaluating a number of languages (e.g. the Web Ontology Language OWL or the Atlas Transformation Language ATL), the Ontology Mapping Language (OML) which was originally developed in the European projects SEKT, DIP and Knowledge Web was selected as most suitable candidate and extended within the HUMBOLDT project to gOML for handling geospatial data. gOML was found to meet the above mentioned requirements. It is very expressive while having a comparatively compact syntax. Mappings are expressed independently of the languages used to describe the data. Thus, gOML can be used with different schema languages like UML or GML application schemas (Reitz et al, 2010b). In the project, the graphical user interface of the HUMBOLDT Alignment Editor (HALE) (see section 3.2) is used to define the mappings which are stored in gOML files. In the course of the EC contract to create the Technical Guidance for the INSPIRE Schema Transformation Network Service, the company 1Spatial has developed a plug-in to HALE (Beare et al, 2010), making it possible to export the mappings to the Rule Interchange Format (RIF), which is recommended in the Technical Guidance document as interchange format for mappings (Howard et al, 2010).

The potential of using ontologies to support automation in schema mapping and to improve the thematic consistency of harmonised data is described by Friis-Christensen et al (2005). However, the authors conclude that further research is needed in this field and that there are still drawbacks. One of the biggest being the lack of formally described ontologies in the geospatial community, partially owed to the fact that special knowledge is required to design ontologies which is also a costly process requiring all stakeholders to agree upon a common domain ontology and provide specific ontologies linked to their application schemas. Similar experiences were made in the HUMBOLDT project where no formal descriptions of ontologies existed for the data used in the HUMBOLDT scenarios.

3.2. Tools and Services for Data Harmonisation

The processing of the mappings from source to target as described above can be achieved in different ways. A variety of approaches, architectures and tools exists today. Architectures can be classified in the following way, taking into account the Draft Implementing Rules for INSPIRE Transformation Services (Drafting Team “Network Services”, 2009, see also Fichtinger and Kutzner, 2010):

1. Offline transformation

In an offline transformation, the Download Service, the Transformation Service, or both are deployed as offline resources. An offline download component can be a file or a local database; an offline transformation component can be locally installed conversion software, where automated workflows can be combined with manual migration steps. Especially for large data sets, very complex schema

transformations and computationally extensive transformation processes like centreline generation or triangulation that have to be performed frequently, network traffic can be reduced by offline transformation.

2. Online transformation

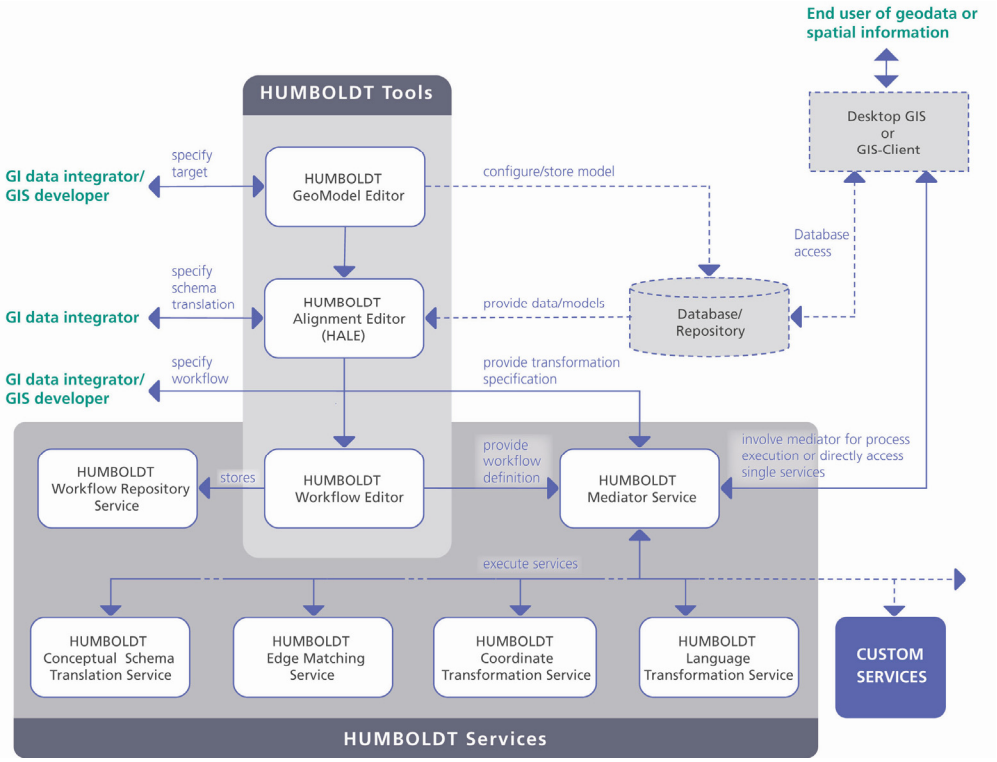
In contrast to offline transformation, online transformations make use of network or web services. Both INSPIRE Download Services and Transformation Services have to be deployed as a web or network service. Regarding the Download Service this means centralised data management and regarding the Transformation Service centralised transformation processing capabilities generic enough to execute different kinds of transformations. The Drafting Team "Network Services" (2009) proposes five different service architectures for Download and Transformations Service. The interaction of the Download and the Transformation Service can be managed by a mediator service as a middleware. Both components can alternatively be incorporated into one single web or network service. On the one hand, the transformation functionality can be encapsulated in the Download Service. In this case, the user directly requests the Download Service. The Download Service performs the transformation and responds with the transformed data. In this case, the transformation component does not necessarily have to follow standards. On the other hand, the Transformation Service can act as a proxy façade. The Transformation Service requests the data for transformation from the download component and responds with the transformed data.

This classification can be further categorised according to the trigger condition for the Transformation Service. The transformation can be performed on the fly or in an asynchronous manner resulting in a harmonised (e.g. INSPIRE compliant) data replication that can be requested from a Download Service component. The transformation can so be performed before publishing the transformed data. This will most likely also be the preferred option for many data providers due to performance aspects.

As discussed in sections 2 and 3.1, heterogeneities in the data are manifested differently. Consequently, in order to deal with such issues, different tools are required. In order to fully use data, data models and the encoding is a valuable input to the harmonisation process. Usually not all data sources used in an application are modelled formally. And when they are, different datasets may have different data models. The HUMBOLDT Framework provides a set of tools that deals each with a single heterogeneity or a group of heterogeneities in the dataset, as illustrated in figure 10. In addition, these tools have been designed and implemented to complement and fill the gap identified in the solutions provided by current state of the art tools. An in-depth description of these tools and how they are applied in a data harmonisation scenario is provided by Fitzner

and Reitz (2009). In this section, the Framework Tools and Services are explained in respect of data harmonisation requirements already discussed previously. Please note that these tools are subject to continued development and that they are not suitable for use in critical production environments.

Figure 10: The HUMBOLDT Framework for Data Harmonisation



Indeed, there already exist a number of tools for data modelling. However, these tools are not specifically designed to be used for geospatial data modelling and thus do not offer all the functionality required in modelling geospatial data. The HUMBOLDT GeoModel Editor is designed and built to enable users to quickly and efficiently deal with geodata modelling tasks. The tool can thus be used for modelling source datasets that do not have schemas or data models as well as for creating the scenarios' common data models. From such models, the tool also allows the automatic generation of XMI and GML schemas.

The HUMBOLDT Alignment Editor (HALE) (Reitz and Kuijper, 2010) is a rich graphical user interface for defining mappings between concepts in GML schemas, as well as for defining transformations between attributes of these

schemas. To make this complex process more accessible to a domain expert and to increase the quality of transformations, HALE allows working with sample instances for visualisation and validation. Furthermore, it includes a sophisticated task-based system that supports users in the creation of a mapping. HALE is not intended to execute the actual transformation. It simply produces a formal representation of the defined transformation, which subsequently has to be processed by a transformation tool or service such as the Conceptual Schema Transformer. In the current status, HALE is optimised for mapping between GML application schemas, but most of the functionality is also available with generic XML schemas.

The harmonisation problems of integrating datasets with different data models are dealt with by the Conceptual Schema Transformation Service (CST). This is a Web Processing Service (WPS) that allows to apply schema transformation to source datasets (mainly feature datasets in GML format) in order to create the target datasets with the target application schema. A schema mapping between the source and the target schema has to be defined in order to accomplish the transformation. These schema mappings are currently defined using HALE and expressed in a high-level language (gOML, see section 3.1). The CST service is described in detail in Reitz et al (2010).

The HUMBOLDT Framework also includes a set of services that are used for processing the data; these are: the Edge Matching Service, The Coordinate Transformation Service and Multiple Representation Merger, Language Transformer Service, and the Workflow Design and Construction Service. These services are further described below.

The Edge Matching Service (EMS) is a WPS implementation of a service that aligns edges and points of vector geometries so that they will be gapless. This tool is used in cases where

1. A data set should be without gaps, but there are small gaps in between the individual features that need to be filled,
2. Datasets that should have identical geometry over a shared feature (such as a common administrative border) have varying geometry,
3. Two datasets having identical geometry for a shared feature have the geometry translated from the position where it should be located.

In the current stage of implementation, it is restricted to the applicability of a simple edge-matching algorithm. More sophisticated algorithms can be integrated as needed.

The Coordinate Transformation Service is a WPS implementation that allows transformation of coordinates between various geographic reference systems.

The Multiple Representation Merging Service is a Web Processing Service that is capable of fusing features of data sets with a spatial overlap, such as along a common border where water bodies are part of both data sets. The Language Transformer is used if data is available in a language different from the one the user wants and is capable of transforming/translating information presented to a user from one language to another based on a multilingual thesaurus.

The HUMBOLDT Workflow Design and Construction Service (WDCS) is composed of two components:

- The Workflow Design Editor which allows the creation of geospatial workflows based on web service technology. Different processing components (in most cases web services that adhere to the OGC WPS specification) can be chained together to realize the required harmonised data output when executed. These composed web services either offer well-known GIS operations such as buffer or intersection calculations, more complex calculations such harmonisation transformations, for instance, coordinate transformation.
- The Workflow Repository Service component is used to store and serve workflows created by using the Workflow Design Editor. This component is accessible via the Workflow Design Editor, and serves workflows in cases where harmonisation of a dataset is needed repeatedly. In which case, the users just query the Workflow Repository to retrieve the previously stored workflow and execute it without having to redesign and recreate the workflow.

The added value of the WDCS component, which makes it different from other workflow tools, is that it provides user assistance for the composition process. For instance, it allows checking the compatibility of the processes in the workflow chain, thus preventing users from connecting two services that cannot be connected due to incompatibility of the inputs and outputs (i.e. when one service processes raster data while the other one processes vector data).

Finally, the Mediator Service executes workflows of processing services and libraries, thus being the central integration component of the framework. It offers clients a number of standard OGC interfaces like Web Map Service (WMS), Web Feature Service (WFS) or Web Coverage Service (WCS); consequently, the component provides interfaces for harmonised data. In contrast to standard download services, it does not hold a data store but assembles the data sets dynamically. This means, in the presence of a service request (OGC GetCapabilities, GetMap, GetFeature etc.), it discovers data sources that either directly match the context of the user (the set of constraints on aspects such as thematic type, spatiotemporal extent or quality elements) or that can be transformed so that they satisfy the context. In case of the latter, it harmonises

the geodata used in the application, both vector and coverage data and ensures that the aggregated data conforms to the user requirements. The Mediator Service provides two main functionalities. First, it handles the actual execution of a transformation defined in a workflow created using the WDCS. Secondly, the Mediator Service provides capabilities useful for many business processes in the HUMBOLDT scenarios through its ability to orchestrate the retrieval, integration and transformation of datasets. Thus, the component allows the realisation of a business process from a client-specific target description. An example of a use case from the HUMBOLDT ERiskA and Atmosphere scenario illustrates this usage in section 3.3.

3.3. Data Harmonisation Put into Practice in the HUMBOLDT Scenarios

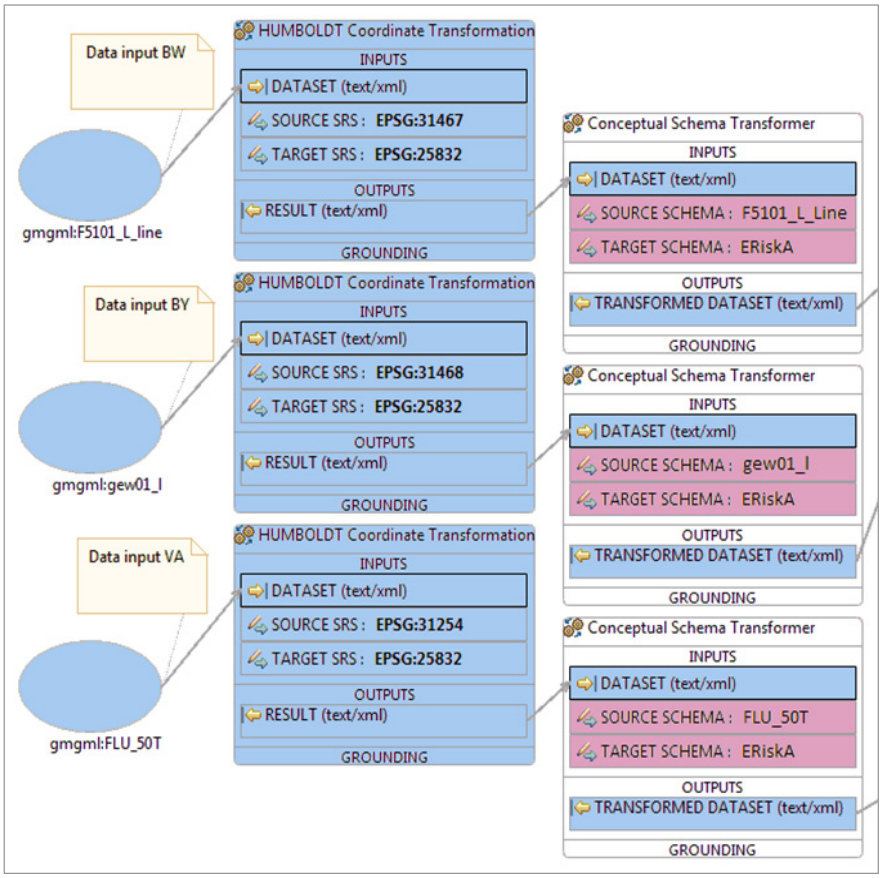
While most of the HUMBOLDT Framework Tools and Services can be used individually, they can also be combined using the WDCS to realise complex geoprocessing tasks as workflows. This is illustrated in the following sections using the HUMBOLDT ERiskA and Atmosphere scenarios (see figures 11 and 12) as examples.

The ERiskA scenario aims at facilitating cross-border cooperation between the agencies responsible for disaster management in the Lake Constance region by enabling the exchange of information on potentially flooded areas and inundation of infrastructure like roads and railways. For this purpose, a number of interoperability and data harmonisation issues are addressed within the Scenario. Similar to the other scenarios, these include, but are not limited to, coordinate reference systems, scales/resolutions, conceptual schemas, metadata profiles and spatial consistency across state/country borders. HUMBOLDT Framework components and tools addressing the harmonisation issues are integrated into the end user application developed for ERiskA to enable the scenario users to access transformation functionalities in a web services based as well as desktop GIS environment.

Figure 11 shows the combination of HUMBOLDT Framework components in a data harmonisation workflow using watercourse data of three states / countries. The workflow is composed using the Workflow Editor. At first, the three datasets of the German states *Bavaria (BY)* and *Baden-Wuerttemberg (BW)* as well as of the Austrian state *Vorarlberg (VA)* are loaded from Web Feature Services (WFS). Each dataset has to be re-projected from its inherent coordinate system to an INSPIRE compliant coordinate system (ETRS89, UTM Zone 32 for the Lake Constance region) by means of the HUMBOLDT Coordinate Transformation Service (CTS). Following the coordinate transformation, each dataset is translated from their legacy schemas to the ERiskA common schema (see Figure 8) by using the Conceptual Schema Transformer (CST) capable of executing complex transformations that have previously been defined in the HUMBOLDT

Alignment Editor (HALE) tool. The three data sets now harmonised according to the ERiskA common schema have to be geometrically aligned using the “Edge Matching Service” (EMS) in the final step. This is required when two semantically equal watercourse features do not meet at the border. At first, the datasets of BY and BW are snapped together with the “Distribute Errors” option, which means that errors in the watercourse geometry are distributed along the border to minimise possible errors. This option is used because both datasets have the same scale (1:25.000). For the harmonisation with the VA dataset, the BY dataset will be set as “Reference Dataset” using the option “Align to Reference” because the VA dataset has a smaller scale 1:50.000).

Figure 11: ERiskA Harmonisation Workflow (detail)



Another HUMBOLDT scenario, Atmosphere, is based on air quality data integration and provision through a Location-based Service (LBS). The scenario

demonstrates possibilities to provide users with air quality information adapted to their needs within a mobile environment. Data used in the scenario are heterogeneous, as they originate from different sources. In each country within the EU, several agencies and local authorities are involved in the collection of air quality data. While the information collected is the same, different data models and encodings are used for data processing. Therefore, the harmonisation of data is necessary before it can be presented to the user.

The HUMBOLDT Framework provides several components to realise this harmonisation. In addition, tools developed within the application scenarios are used for processing the harmonised data. One of these processes is the generation of coverages from the measurement data of the air quality components. This dataset is provided by the European Environmental Agency (EEA) in a flat xml format. Before the data is provisioned to the user, it needs to be processed.

Within the scenario application, two main harmonisation issues are encountered. First, because the data model in the source dataset differs from the target application schema, schema harmonisation is required. Second, point data needs to be merged with coverage data, to enable the retrieval of air quality information from any location.

In a scheduled process, source datasets are downloaded every 30 minutes, parsed and stored in a spatially enabled relational database. This facilitates the direct use of the database table as a data-store by Geoserver WFS. This action is achieved by a component that has been developed as part of the scenario demonstrator.

At first, the HUMBOLDT Alignment Editor is used to define the mapping from the source schema to the target schema. Once such mappings have been defined, the Conceptual Schema Transformation Service is used to execute the actual transformation of data from the source to the target schema.

The Atmosphere Scenario furthermore gives us an example for the need of a Workflow Repository Service (WRS) allowing adapting predefined and reusable workflows to concrete user requirements. Since users need to retrieve air quality data irrespective of their location, the point data must be interpolated to generate a coverage dataset of air quality information. An interpolation service is used to perform this task. The service takes a dataset (air quality measurement data in GML format) as input and interpolates the values for each component in the input dataset, i.e. the point values from all the stations for each component in the data are interpolated to create GeoTiff coverage. As a service-specific requirement the algorithm used by the service only accepts dataset input in a specific coordinate system (UTM coordinate system). However, the source data is generated with

geometry in geographic coordinate systems (EPSG:4326). This requires coordinate transformation before the dataset can be interpolated. For this purpose, the Coordinate Transformation Service WPS is used. The resultant GML is then parsed to generate a comma-separated file (CSV with coordinates and measurement values) which is used as an input by the interpolation service. The resulting coverages are automatically published to a Web Coverage Service (WCS) on the Geoserver. This requires a pre-configuration of a coverage-store and coverage for each air quality component in the dataset.

All this back-end pre-possessing is achieved using an array of processing/harmonisation services offered by the Framework and combined together using the WDCS. All data requests from the LBS application server are directed to the Mediator component which executes the predefined workflow for data harmonisation and processing. This ensures that the data can satisfy the end users' requirements.

4. CONCLUSIONS AND OUTLOOK

Data harmonisation is not a new problem. It has been around for as long as heterogeneous data has to be combined in applications. A lot of solutions have been developed in the general IT world. To solve the specific harmonisation issues in the geospatial world, general IT solutions have been adapted and specific "spatial" solutions have been developed. Currently, there are a number of tools available to solve individual steps in the data harmonisation process. Yet, there is no open source tool that covers and integrates all process steps at this time. Furthermore, existing tools can often not be easily integrated into service-oriented architectures. One of the major goals of the HUMBOLDT Framework therefore was to facilitate data harmonisation as part of the overall processing of an information request. The HUMBOLDT Framework is designed to be minimally invasive, i.e. not to replace existing systems but rather support and amend them with specific capabilities needed in the data harmonisation process. Therefore, the functionality of the HUMBOLDT Framework is well-isolated from the interfaces by which they are accessed, resulting in components that can be easily adapted and extended for different deployments and process synchronisation styles. The data harmonisation process is performed flexibly and, partially, depending on the workflow, automated in several different steps. This flexibility and partial automation are the main benefits of the HUMBOLDT Framework. It can provide users with solutions tailored to their respective requirements and existing infrastructure. Targeted user communities include government agencies at different administration levels, cross-border projects, research projects, or citizens' initiatives.

The INSPIRE directive has added additional momentum to the development of data harmonisation tools in the geospatial domain, also taking into consideration

aspects like service-oriented architectures. However, there is still a lack of commonly agreed standards in the geospatial community, e.g. for languages to formalise the mapping between two schemas. The methodology, rules and guidelines developed by INSPIRE to facilitate the creation of the European Spatial Data Infrastructure have been tested in the HUMBOLDT project. They have proven to be useful, e.g. for the developed cross-border applications. However, it has become apparent during the project that the model-driven approach, e.g. conceptual modelling using UML, is fairly new to many actors in the geospatial domain and that there often is a lack of formalised representations of existing data structures in conceptual schemas. The information analysis in the different scenarios also revealed a multitude of heterogeneities. Some can be solved by rather simple transformations, others require very complex transformations or even might not be solvable at all, e.g. when the semantics of concepts are too different. The results of testing INSPIRE Data Specifications in HUMBOLDT can be summarised as follows: in many cases it was possible to create mappings from the schemas of the legacy data to the INSPIRE schemas, but often information is missing in the legacy data sets in different countries to fill all the attributes in the target schema. In some cases several data sets, sometimes even from different agencies, will have to be combined. This may be the case e.g. for the hydrography example in this paper. The topographic data used in the scenario has been created by mapping agencies mainly for spatial reference and mapping purposes. For use in sophisticated hydrological modelling or also environmental monitoring and reporting as foreseen in one of the use case in the INSPIRE Data Specification on Hydrography, further information e.g. held by specialist (e.g. water management agencies) agencies is needed.

The HUMBOLDT results, i.e. the software framework for data harmonisation and service integration, the data harmonisation and INSPIRE testing experiences from the different scenarios as well as the training programme, can provide valuable contributions to the implementation of INSPIRE and SDIs at different levels and can help opening up new fields of application for spatial information. Therefore, to gain as much benefit as possible, the HUMBOLDT Open Source Framework and its toolset are available to all. Further development for the community is a major goal for the HUMBOLDT consortium. This also includes additional services as well as adaptation to upcoming requirements and new application areas.

The experience gained during the development of HUMBOLDT Tools and Services does not only embrace technically related issues but also has a strong focus on data harmonisation issues in general. This expertise in data harmonisation processes and methods as well as their implementation in HUMBOLDT Tools and Services is of great interest and importance for the GI community, and can be effectively used to develop various new products and

service offerings related to the HUMBOLDT Framework and data harmonisation processes and methods in general.

HUMBOLDT Tools and Services are furthermore of great potential interest for future European projects related to INSPIRE or other projects where the needs of interoperability and data harmonisation as well as standardisation will be of importance. The exploitation of the HUMBOLDT project results in research will focus on future European research or educational projects, as well as on the running ones, such as eSDI-NET+, GENESIS, GS-Soil, ESDIN, Plan4all, VESTA-GIS, NatureSDIplus, BRISEIDE, etc.

The development of a long-term sustainability concept for HUMBOLDT addresses the adoption, implementation, use and continued support of the system by resource-holders, including government authorities, technical personnel and users. The selection of an Open Source strategy alone does not guarantee sustainability by itself. To sustain the HUMBOLDT Framework as Open Source solution implied that the project had to grow not only as an isolated framework, but also to design ways of making the framework easily accessible to other developers and users. Since March 2009, the HUMBOLDT project is open to external communities through a forum, wiki and externally accessible code repositories. The continuation of the HUMBOLDT Community Website as well as the continued development, bug fixing and maintenance of the HUMBOLDT components are of major importance for the sustainability and post-project exploitation of the HUMBOLDT Framework and will be secured by the HUMBOLDT consortium.

Interested readers can access additional information on HUMBOLDT on the training platform which offers specific training packages that are addressing both users and software developers. These training opportunities not only embrace the user communities of the HUMBOLDT Tools and Services but also the communities of GMES and INSPIRE.

In summary, it can be stated that the tools and the methodology for data harmonisation developed in the HUMBOLDT project can be of major importance in various fields. Besides data harmonisation related issues across various application fields, there is great potential for the HUMBOLDT results to be exploited for education, training and research on a local, national, regional and international levels.

ACKNOWLEDGEMENTS

This paper describes work carried out by the partners of the HUMBOLDT project which is funded under the Sixth Framework Programme of the European Union

(Aerospace / GMES) for the period 01/10/2006 – 31/03/2011, with contract number SFP5-CT-2006-030962.

REFERENCES

Beare, M., Payne, S., and R. Sunderland (2010). Prototype Report for the INSPIRE Schema Transformation Network Service, V. 3.0, at http://inspire.jrc.ec.europa.eu/documents/Network_Services/JRC_INSPIRE-TransformService_ProtoRpt_v3-0.pdf, [accessed 16 December 2010].

Commission of the European Communities (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 (INSPIRE), *Official Journal,* L 108, 25/04/2007, pp. 1-14.

de Vries, M., Giger, Ch. and M. Loidold (2007a). *A3.5-D1 State of the Art in Data Harmonisation and Data Management*. Darmstadt: HUMBOLDT Consortium.

de Vries, M., di Donato, P. and F. Penninga (2007b). *A7.1-D1 Concept of application-specific harmonised data models*, Darmstadt: HUMBOLDT Consortium, at http://www.esdi-humboldt.eu/files/818-a7_1d1_concept_of_application-specific_harmonised-tud-001-final.pdf, [accessed 7 June 2010].

Donaubauer, A., Straub, F. and M. Schilcher (2007). “mdWFS: A Concept of Web-enabling Semantic Transformation”. *Proceedings 10th AGILE Conference on Geographic Information Science, 8 - 11 May 2007, Aalborg, Denmark*, at http://people.plan.aau.dk/~enc/AGILE2007/PDF/157_PDF.pdf, [accessed 10 June 2010].

Drafting Team "Data Specifications" (2008). *Methodology for the development of data specifications,* at http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/D2.6_v3.0.pdf, [accessed 2 June 2001].

Drafting Team "Data Specifications" (2009). *INSPIRE Generic Conceptual Model, Version 3.2*, at http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/D2.5_v3.2.pdf, [accessed 2 June 2010].

Drafting Team “Network Services” (2009). *Draft Implementing Rules for INSPIRE Transformation Services*, at http://inspire.jrc.ec.europa.eu/documents/Network_Services/INSPIRE_Draft_Implementing_Rules_Transformation_Services_%28version_3.0%29.pdf, [accessed 15 June 2010].

Fichtinger, A. and T. Kutzner (2010). “Datenharmonisierung im Kontext von INSPIRE”, in M. Schilcher (Ed.). *Tagungsband 15. Münchner*

Fortbildungsseminar Geoinformationssysteme. Heidelberg: abcverlag, pp. 222-238.

Fitzner, D. and Th. Reitz (2009). *A5.2-D3 [3.0] A Lightweight Introduction to the HUMBOLDT Framework V3.0*. Darmstadt: HUMBOLDT Consortium.

Friis-Christensen A., Schade S. and S. Peedell (2005). "Approaches to solve schema heterogeneity at European Level", *Proceeding 11th EC-GI & GIS Workshop, ESDI: Setting the Framework, 29 June – 1 July 2005, Alghero, Sardinia, Italy*, at http://www.ec-gis.org/Workshops/11ec-gis/papers/301peedell.pdf, [accessed 11 June 2010].

Howard, M., Payne, S. and R. Sunderland (2010). *Technical Guidance for the INSPIRE Schema Transformation Network Service*, V. 3.0, at http://inspire.jrc.ec.europa.eu/documents/Network_Services/JRC_INSPIRE-TransformService_TG_v3-0.pdf, [accessed 16 December 2010].

Lehto, L. (2007). "Schema Translations in a Web Service Based SDI", *Proceedings 10th AGILE Conference on Geographic Information Science, 08-11 May 2007, Aalborg, Denmark*, at http://people.plan.aau.dk/~enc/AGILE2007/PDF/29_PDF.pdf, [accessed 3 June 2010].

Reitz, Th. and A. Kuijper (2010). "Applying Instance Visualisation and Conceptual Schema Mapping for Geodata Harmonisation". *Advances in GIScience: Proceedings of the 12th AGILE Conference.* Berlin, Heidelberg: Springer, pp. 173-194.

Reitz, Th., Schäffler, U., Klien, E., Fitzner, D. (2010). "Efficient Conceptual Schema Translation for Geographic Vector Data Sets", *Proceedings 13th AGILE International Conference on Geographic Information Science, 10 - 14 May 2010, Guimarães, Portugal*, at http://agile2010.dsi.uminho.pt/pen/ShortPapers_PDF%5C112_DOC.pdf, [accessed 11 June 2010].

RISE (2007). *Methodology & Guidelines on Use Case & Schema Development Version 1.2.*, at http://www.eurogeographics.org/documents/RISE15MethodologyandGuidelinesV1.2.pdf, [accessed 11 June 2010].