HUMBOLDT PROJECT FOR DATA HARMONISATION IN THE FRAMEWORK OF GMES AND ESDI: INTRODUCTION AND EARLY ACHIEVEMENTS

P. Villa[a], T. Reitz[b], M. A. Gomarasca[a]

[a]CNR-IREA, Institute for Electromagnetic Sensing of the Environment, Via Bassini 15, Milano, Italy - villa.p@irea.cnr.it
[b]Fraunhofer-IGD, Institut Graphische Datenverarbeitung, Fraunhoferstraße 5, Darmstadt, Germany

Commission VIII, WG VIII/2

KEY WORDS: Geoinformation, SDI, GMES, INSPIRE, Architecture, Data Harmonisation

ABSTRACT:

The issue of spatial information sharing and harmonisation has become of the uttermost importance in the European context, with the European Community strongly engaged not only in political and economic changes and integration, but more and more tackling technical and scientific problems. In the field of geospatial information, the need for harmonised and interoperable data is now a key topic both for users and producers of geodata. This has brought the initiatives related to the INSPIRE Directive, recently entered into force and soon to be receipted by national legislative frameworks, and to the GMES programme, now entered into its operational phase, both of them oriented to enable cross-border and cross-sector usage of geoinformation, aiming at the realization of the future European Spatial Data Infrastructure (ESDI).The HUMBOLDT project, started in October 2006, is supported by the European Community through the 6th Framework Programme and has the aim of bring together a variety of scientific, technical, economic and policy driven points of view with the aim of implementing a Framework for harmonisation of data and services in geoinformation domain.The two-pronged approach of HUMBOLDT comprises a technical side of framework development and an application side of scenario testing and validation, through an iterative refinement of the harmonisation solutions provided within the project. The Architecture of the HUMBOLDT Framework has been based on an approach which comprises as the fundamental part a Mediator Proxy able to support standard interfaces, side by side with specialized interfaces and also supporting the description of transformation rules from the viewpoint of the conceptual schema level. All those aspects and issues are being tackled within the HUMBOLDT project environment, and will be the core of the project until its official ending and beyond, towards the establishment of an ESDI.

1. INTRODUCTION

With the ascension of Bulgaria and Romania, there are currently 27 Member States in the European Union. While the political and economic integration process makes good progress, the topic of geoinformation has traditionally been scattered and fragmented, even within single countries (Annoni and Smits, 2003).

Spatial information has been available for centuries in form of paper maps. Since no trans-national coordination has been given, a scattered set of spatial data evolved which hampers and partly prevents the exchange of spatial information between countries. The lack of cross-border solutions becomes obvious for instance in case of disasters which do not stop at national borderlines.

Right now the European Community faces the growing need for available, reliable and interoperable geo-data, in the frame of the establishment of the future European Spatial Data Infrastructure (Smits and Friis-Christensen, 2007). The need for harmonised geoinformation is therefore becoming a key topic for geo-data users and producers and thus also in geosciences (Bernard and Craglia, 2005). This has also been a driver or the INSPIRE initiative and the GMES programme, both initialized by the European Commission and enabling cross-border and cross-sector usage of geoinformation over an European Spatial Data Infrastructure (ESDI).

In an ideal world, geographic information would be harmonised and interoperable. Harmonisation refers to the standardization of data so that they can be matched with other data and information regardless of the format. Interoperability is the ability of products, systems, or business processes to work together to accomplish a common task. In terms of standards, many international organizations are working, like ISO and OGC. Another important aspect is related to the data itself. When we speak of data we mean data and metadata at the same time. Metadata is 'data about data'. Although metadata creation might seem quite logical and inherent to the production of datasets, especially regarding geographical datasets, lack of metadata remains one of the main problems coastal managers face frequently.

The major issue in data harmonisation is that currently there is no actual framework setting common, unifying, data collection and production measures and a proper agreed exchange format between organisations. This state of affairs has been the trigger for the development of the European spatial information infrastructure initiative which has led to the INSPIRE Directive.

2. THE HUMBOLDT PROJECT

The HUMBOLDT project was started by October 2006 and will run for 4 years. Under the coordination of Fraunhofer Institute for Computer Graphics 28 project partners from 14 European Countries work together on 12 work packages. The project partners represent private companies, public authorities, universities and other research institutions, which prevents the

project from taking any kind of narrow perception of spatial data and processes.
The work packages are organized along a two-pronged approach (see Figure 1) has been designed and followed. This approach focuses on integrating both concrete application requirements but also technical innovations, best practices and research results.

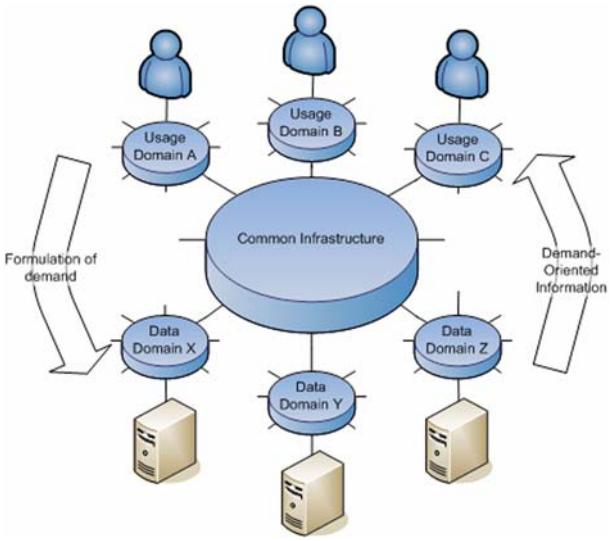

Figure 1. The user-driven approach of HUMBOLDT project to geodata harmonisation

An essential element of the application driver side of the project is the development of so-called HUMBOLDT scenarios in which the different components are applied and tested under realistic conditions and which represent different GMES application fields, ranging from border security to urban planning, risk management and the protection of nature reserves. For these scenarios, a process analysis which shows the steps necessary to harmonise data and metadata has been conducted for two scenarios and is currently being conducted for another six scenarios. The different stakeholder groups have to act as drivers, to ensure that the technological development fulfils their requirements as good as possible.

This analysis gives the base for the specification and development of a software framework and diverse tools to be integrated into the ESDI to support spatial data and service providers in offering standardized spatial information.

Also, influences from the technical side, such as the state of the art in research and development as well as in standardisation are identified, organized and tracked. The result of these influences is the framework development, i.e. the production of a set of software components and tools that enables geoinformation users and specialists to link their resources and processes as seamless and effortless as possible into the evolving spatial data infrastructure.

At the core of the development work done within the HUMBOLDT project stands therefore the HUMBOLDT framework. This software framework is the hull for the various data harmonisation scenarios and provides the common functionality required by the scenarios. The framework itself as a project outcome mainly addresses the developer community and the application developers using it to create end-user applications and scenarios.

Besides offering a base for scenarios and project-external data harmonisation projects, the framework will also profit from this usage and the feedback that will inevitably be generated. By these scenarios and projects, the framework will be evaluated, and it will be tested whether it is developer-friendly, addresses user's needs and it robust enough to scale and fit to the individual business requirements.

The HUMBOLDT project contains a multitude of research activities that have been grouped to challenges (Villa *et al.*, 2007). Among these challenges are to identify user requirements, the re-use of existing results, maximizing the usefulness of spatial data, to create the software in such a distributed group, and to allow an efficient use of harmonized geoinformation. The most important of the topics addressed in these challenges are:

1. Identification and description of users and their typical geoinformation harmonization requirements
2. Collection of the State of the Art and evaluation of existing or developing Standards, both to ensure not missing developments and implementation experiences
3. Identification of cross-cutting concerns of an ESDI, such as security and access control, infrastructure concerns such as supported interfaces, load distribution and redundancy and also of functional concerns
4. Easily understandable visualizations for complex data
5. Development of a quality measurement approach for semantic mappings
6. Creation of a method to express trust in semantic mappings
7. Automatic transformation of geoinformation and metadata using semantic mappings using a partly visual harmonization application
8. Development of methodologies to maintain consistency between multiple representations from different sources
9. Efficient integration of legacy services
10. DRM and access control for geoinformation
11. Efficient and parallel processing of geoinformation
12. Adaptable interfaces such as streaming to handle different types of clients and different modes of processing
13. Orchestration of Geoservices, especially finding a way on how to orchestrate over different types of Web Service interfaces such as SOAP/WSDL, OGC-WS or REST-style web services
14. Creation of data profiles for the geoinformation required in the scenarios based on the common data models provided by INSPIRE
15. Creation of an evolutionary specification and implementation process for agile, distributed development
16. Development of best practices and patterns for ESDI applications.

Currently, a first version of the HUMBOLDT Framework is available, together with a Grounding Catalogue and an initial set of data for the assessment of these prototype functionalities with special reference to application areas indicated by HUMBOLDT Scenarios.

To summarize the potential impact, the benefits of HUMBOLDT can be beneficial to a wide range of organizations, from political institutions to scientific research to commercial enterprises, reaching at the end of the path the European citizen, more and more an aware user of geoinformation data and services, soon to be included in the frame of the ESDI.

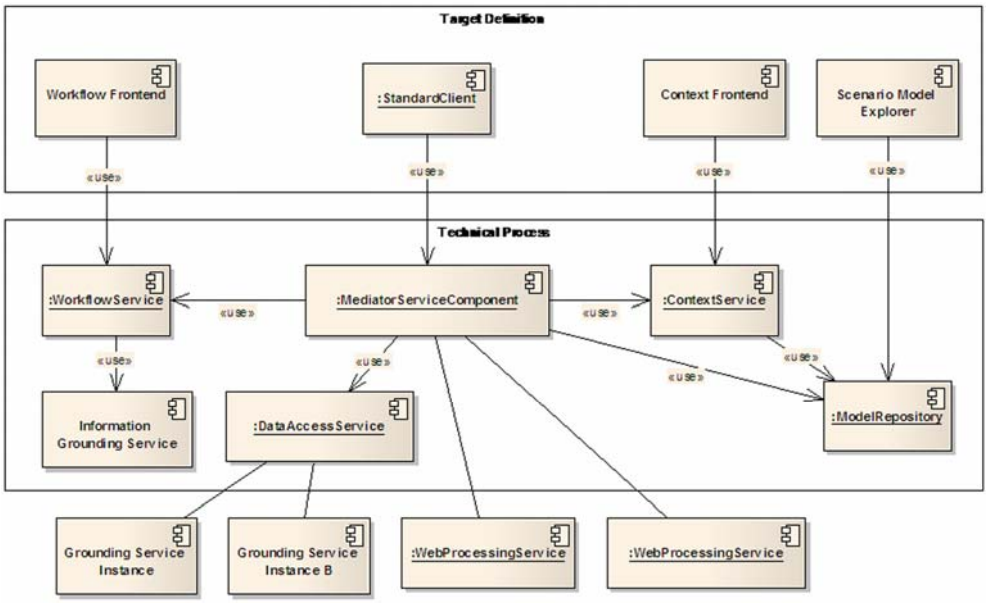

Figure 2. Overview of HUMBOLDT Framework Components

3. ARCHITECTURE FOR A USER- AND TASK-DRIVEN SDI

After initial analysis of the wide range of harmonization requirements and processes in the HUMBOLDT scenarios as introduced above, we found that even though certain patterns occur between the requirements as indicated in Table 1, there are highly diverse environments and processes to integrate with. Consequently, we have taken a step back from viewing our main business process as being the one-way harmonisation process and tried to get a wider view on geodata integration issues.

INSPIRE Harmonisation Requirement	Scenario Priority Points
Application schemas and feature catalogues	16
Coordinate referencing and units	15
Portrayal Model	6
Metadata Schema	17
Metadata Handling	13
Data Transfer (Encoding)	11
Data capturing	3
Terminology	15
Spatial and temporal aspects	12
Multiple representations	16
Object referencing modelling	6
Identifier Management	14
Maintenance	5
Consistency	7
Conformance	6
Multi-Linguality	7
Data translation model	5
Registers and registries	9
Quality	17

Table 1. Priority Points for different Data Harmonisation aspects, based on an expert analysis of seven HUMBOLDT scenario specifications (importance per scenario valued from 0 to 3, maximum SPP is thus 21)

This view of an abstract integration business process which encompasses very different scenarios has also been guiding in the creation of the Framework Architecture. Consequently, one of the core goals was not just to facilitate integration processes of non-standardized data sets into the ESDI but also to provide users with an adaptive view on the SDI resources. A central aspect to this is the modelling of the context of the SDI, the user's organization, the user himself and his task. In total, these context elements define all requirements a user has for accomplishing a certain task. From this context, a product description can be inferred which is then used to filter down all resources available to those which are relevant. The step of discovering data sets or services is thus done automatically instead of manually by browsing through a catalogue.

One specific class of tasks we are focusing our attention on is the process of integrating geodata into a spatial data infrastructure. For this, the geodata has to be transformed in many aspects ranging from spatial reference systems to language, encoding and classification. Consequently, the context of this geodata harmonisation task are the rules imposed by the SDI into which to publish geodata; in the case of the ESDI described by INSPIRE this would consequently be the implementation rules (IR) currently being created by the INSPIRE Drafting Teams (Portele *et al*., 2007). This approach has the advantage of being very flexible and allowing a wide range of different transformation targets.

Another guiding principle of the architectural work in HUMBOLDT was not just to hide SDI complexity and information masses from the user where they would be unnecessary anyhow, but also to hide the complexity of the data harmonisation taking place. This is especially of interest in those processes where a user is not publishing data into an infrastructure, but rather wants to integrate data from the infrastructure for use in his local, domain-specific applications (Reitz *et al*., 2006). To fulfil this principle, we have taken to the well-known approach of a Mediator Proxy, which offers only standardized established interfaces like OGC WMS, WFS or

WPS. This allows our users to continue working with the tools (and relatively simple interfaces) they are used to work with.
However, there is a second layer of interfaces, both for humans and for machines, which are more specific to defining integration processes and goals. Among those are:

- A Workflow Definition and Construction Service which stores abstract workflows and also handles their expansion to concrete executable workflows depending on the user's set of context constraints, which are provided to the Mediator Proxy for execution;
- A Context Service, which provides the Mediator Proxy with all information of a given request's user, organization and SDI imposed rules;
- A Model and Mapping Repository, which provides the Mediator Proxy with formalized conceptual schemas and executable transformation descriptions between different conceptual schemas;
- An Information Grounding Service, which relates concepts from the conceptual schemas stored in the Model Repository to concrete services.

Furthermore, there is a set of so-called Transformer Services, which are essentially Web Processing Services offering a few additional capabilities such as asynchronous communication. These provide capabilities for actual transformation of geodata sets ranging from edge matching to language and conceptual schema translation. Again, leveraging a standardized interface, we can make use of processing services developed outside the project. A general overview of the HUMBOLDT framework Components is given in Figure 2.

The two processes described so far (integration of data into an infrastructure and usage of infrastructure data in a specific domain) cannot however be automated without substantial effort beforehand. With the solutions in place in our scenarios, high effort has to be spent annotating existing services and creating conceptual schemas for those. Also, the description of transformation rules especially on the conceptual schema level is a very hard task for domain experts (from fields such as protected area management, remote sensing and oceanography), as a recent user study conducted within the project has shown. Consequently, our third main business process, which actually has to happen before any transformations can be executed, is the source and target schema description and publishing as well as the description and publishing of the schema mapping connecting the two. Furthermore, the abstract workflows for each task class (such as *Create Map*, *Retrieve Capabilities*, or more specific tasks like *DetermineAffectedAreas*) have to be created. These are however re-useable in wide range of applications, since they are tailored towards a certain task only when being transformed into executable workflows.

As an example for this re-use in the case of *DetermineAffectedAreas*, consider an oil spill in the Channel between Great Britain and France (see Figure 3). The oil spill endangers protected areas in the channel, but also economy like oyster farms. To find out what exactly is endangered, an operator would normally access several data sets and see how the simulated drift of the oil spill will spatially coincide with those. Using our approach, the operator has a concept of *SusceptibleArea* in his application schema which has subconcepts like *NatureReserve*, *SusceptibleStructure* and *Aquaculture* areas. Now, the operator only needs to invoke the task *DetermineAffectedAreas* with the concept *SusceptibleAreas* and the output from the drift simulation run. Using some additional constraints like the time for which susceptible areas are to be determined, the operator is provided with a data set or portrayed product containing the information required. This very same process, used with different concepts from a different application schema, will be useful to an operator from a different domain, such as an environmental expert working on finding out what impact a new road will have on an ecosystem. This is also facilitated by abstracting concrete processing implementations and tasks from each other.

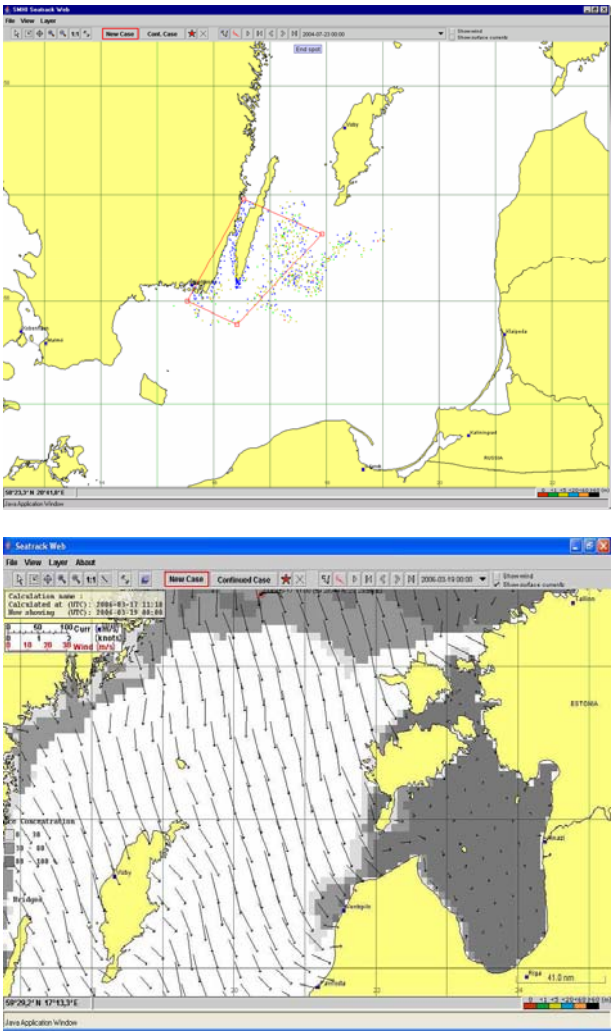

Figure 3. An example of Graphical User Interface for HUMBOLDT Scenario covering Ocean geodata for oil spills monitoring, integrated with SMHI Seatrack Web service. Data from various sources are collected and processed to map the oil spill depth forecasting (upper image) and to foresee the future position of the oil, through the integration of sea current data (lower image).

For all those activities in this third main business process, several specialized tools are being developed in on-going research in the project, among them a tool for application schema design, a tool for specifying transformation processes and a tool for specifying conceptual schema alignments.

To summarize, we have created an architecture which can be integrated in a wide range of different environments, that covers

a wide range of data integration tasks and that is accessed both via well-known interfaces for most users and via specialized and powerful user interfaces for domain experts.

4. SUMMARY AND OUTLOOK

The European Community faces the growing need for available, reliable and interoperable geo-data, in the frame of the establishment of the future European Spatial Data Infrastructure. The need for harmonised geoinformation is therefore becoming a key topic in geosciences and geo-data users and producers (Bernard *et al*., 2005). Putting INSPIRE principles into practice and following the developments of GMES, the HUMBOLDT project has the goal of supporting and advancing the process of definition and implementation of the ESDI, by providing a software framework for geo-data harmonisation and geo-services integration.

The two-pronged approach of HUMBOLDT comprises a technical side of framework development and an application side of scenario testing and validation, through an iterative refinement of the harmonisation solutions provided within the project.

The Architecture of the HUMBOLDT Framework has been based on an approach which comprises as the fundamental part a Mediator Proxy able to support standard interfaces. More specialized interfaces are also integrated in the Framework Architecture, as:

- A Workflow Definition and Construction Service;
- A Context Service;
- A Model and Mapping Repository;
- An Information Grounding Service;
- A set of Transformer Services.

The transformation and harmonisation process and its feasibility and efficiency, strongly depends on the availability of the description of transformation rules from the viewpoint of the conceptual schema level.

All those aspects and issues are being tackled within the HUMBOLDT project environment, and will be the core of the project from now until 2010, which is the official end of the project, and beyond, on the path leading to the actualization of INSPIRE efforts into an ESDI.

The research so far has indicated very well where our efforts have to be focused for the next phase of the project. Among the concrete items currently investigated are:

- Deeper investigation of the process of geoinformation harmonization itself on a larger basis, to see what patterns emerge between data providers and consumers;
- Integral quality management in all data processing steps;
- Handling very different types of data in one contiguous infrastructure;
- Efficient execution of the process, taking into account optimizations possible by combining multiple steps;
- Actual execution of conceptual schema translations
- Development of user guidance concepts for geospatial data alignment processes, including development of languages for geospatial domain experts to describe alignments;
- Handling of semantic mismatches.

REFERENCES

Annoni, A., Smits, P.C., 2003. Main Problems in Building European Environmental Spatial Data. *International Journal of Remote Sensing*, vol. 24, no. 20, pp.3887-3902.

Bernard, L., Craglia, M., 2005. SDI – From Spatial Data Infrastructure to Service Driven Infrastructure. *1st Research Workshop on Cross-learning on Spatial Data Infrastructures (SDI) and Information Infrastructures (II)*, Enschede (The Netherlands).

Bernard, L., Kanellopoulos, I., Annoni, A. & Smits, P.C., 2005. The European Geoportal – One Step Towards the Establishment of a European Spatial Data Infrastructure. *Computers, Environment and Urban Systems*, vol. 29, pp. 15-31.

HUMBOLDT Project, 2008. Towards the Harmonisation of Spatial Information in Europe, project public website: http://www.esdi-humboldt.org.

Portele, C., Van Oosterom, P., Bayers, E. *et al*., 2007. INSPIRE Methodology for the development of data specifications. *INSPIRE Drafting Team "Data Specifications"* (May 2007).

Reitz, T., Holweg, D., Ludlow, D. *et al*., 2006. HUMBOLDT – Development of a framework for data harmonisation and service integration – Description of Work; *Annex I of the HUMBOLDT project contract* (October 2006).

Smits, P.C., Friis-Christensen, A., 2007. Resource Discovery in a European Spatial Data Infrastructure. *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, pp. 85-95.

Villa, P., Reitz, T., Gomarasca, M.A., 2007. HUMBOLDT project: implementing a framework for geo-spatial data harmonization and moving towards an ESDI. *Geoinformation in Europe, Proceedings of the 27th EARSeL Symposium*, Millpress, pp. 29-36.