GeoInformation in Europe, M.A. Gomarsca (ed.)

The HUMBOLDT project: implementing a framework for geo-spatial data harmonization and moving towards an ESDI

Paolo Villa
IREA-CNR, *Institute for Electromagnetic Sensing of the Environment, Via Bassini 15, Milan, Italy, email: villa.p@irea.cnr.it*

Thorsten Reitz
Fraunhofer-IGD, *Institut Graphische Datenverarbeitung, Fraunhoferstraße 5, Darmstadt, Germany*

Mario A. Gomarasca
IREA-CNR, *Institute for Electromagnetic Sensing of the Environment, Via Bassini 15, Milan, Italy*

Keywords: Data harmonisation, Spatial Data Infrastructure, INSPIRE, Geo-services integration, interoperability, GMES

ABSTRACT: 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. 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 collaboration with other geoinformation projects, the integration of a Review and Advisory Board, the establishment of both a developer and user communities are among the project components that are intended to conduct the efforts of the HUMBOLDT consortium and focus the work towards efficient, reliable and economic solutions to the challenges of data harmonisation throughout Europe.

The benefits of HUMBOLDT will range 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.

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.

As an example for this, there are 15 different tide gauge reference points in use in the European Union. The differences between National Vertical Datums and UELN 95/98 (United European Levelling Network) vary from −231 cm (for Belgium) to +22 cm (for Finland). Thus, actual interoperability in all kinds of application areas affected by any geodetic reference system, such as emergency response in flood events, is heavily obstructed by trans-national reference system discrepancies (Annoni & Smits 2003). Soon to become active ESA satellite programs Galileo and GOCE will provide new insights and possibilities to homogenize the vertical reference system, but once again the problem of data harmonization is only moved from reference system to satellite data interoperability (Marchetti 2007).

Vertical Datum inconsistencies are only one example of the issues posed to spatial information sciences in the field of data heterogeneity; data format, coverage, scale, reference system, data model, ontologies and metadata schemata are all affecting the limits of application of geo-spatial data. Considered all together, those discrepancies raise the strong need for international level harmonization initiatives in the field of geoinformation.

At European level, the Commission – mainly through the work of the Joint Research Centre – has therefore envisaged a proposal for a Directive regarding the establishment of an infrastructure for spatial information in the Community, briefly called INSIPRE (which stands for INfrastructure for SPatial InfoRmation in Europe), recently adopted by the European Parliament and entered into force in May 2007.

AS a consequence of this, during the next few years the European Union will establish a European Spatial Data Infrastructure (ESDI), and one of the main efforts in implementing this infrastructure is the challenge of geo-data harmonization throughout Europe (Annoni & Smits 2003, Smits & Friis-Christensen 2007). A number of international projects funded by the European Community are now dealing with the issues and difficulties described above, among which HUMBOLDT has the objective to provide an effective framework for data harmonization and service integration, with the general aim of making easier for European geospatial information users to exploit geoinformation knowledge in new ways (Reitz *et al.* 2006).

2 HARMONISATION: THE NEEDS FOR HUMBOLDT

The need for harmonized data is a fundamental point in building a Spatial Data Infrastructure which comprises different data sources and foresees different services and applications for retrieved geo-data (Annoni & Smits 2003, Bernard & Craglia 2005, Toth & Nunes de Lima 2005).

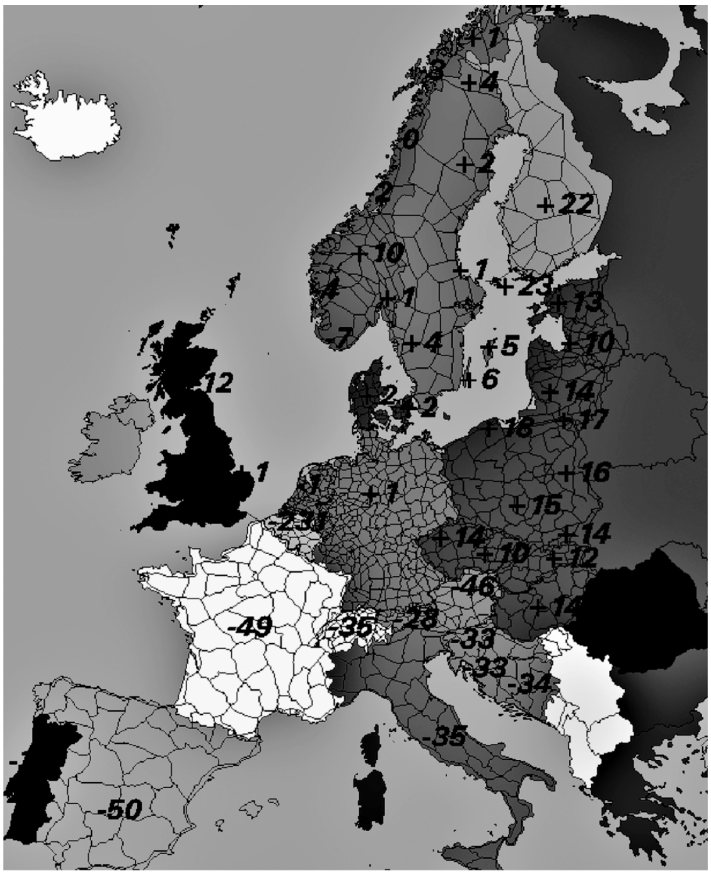

Figure 1. Height differences between National Vertical Datums and UELN 95/98 for most of the European Countries, measured and shown in centimetres.

According to INSPIRE guidelines, the structure of an ESDI shall be composed of a set of interoperable, interacting services, thus following the Service Oriented Architecture (SOA) paradigm. Such an architecture matches the distributed responsibilities regarding service provision and data management in the geoinformation sector well. For a SOA to work, an essential element is to select or build a group of interface standards that are mutually interoperable and complementary.

The widely used interfaces which largely fulfill this requirement are the Open Geospatial Consortium (OGC) Web Services, specifically the Web Map Service (WMS), the Web Feature Service (WFS) and also the Catalogue Service Web (CSW), as a means to publish, find and use services. Any new component for SDI should be interoperable with these existing services and at the same time must be able to adapt to future developments. Besides these well-established standards, there are various areas where standardization has not come very far yet or where there are multiple competing standards. HUMBOLDT

developments must therefore be very flexible with respect to their configuration and modes of deployment to fit into existing spatial data infrastructures.

Within the scope of the HUMBOLDT project, after a first analysis of harmonization-related studies and efforts, the following working definition for data harmonization is taken into account:

Geo-data harmonization implies and means 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 Geo-data harmonisation implies and means the possibility to combine data from heterogeneous sources into seamlessly integrated, consistent and unambiguous information products, in an easy and repeatable way, adapted to the end-user's requirements and context (De Vries *et al.* 2007).

In practice, several reasons and kinds of heterogeneity are present in the geoinformation field; a number of classifications of heterogeneity are possible. Often a distinction is made between syntax, structure and semantics (Friis-Christensen *et al.* 2005):

- Syntax (related to different data formats, e.g., db, shape files or MapInfo),
- Structure (related to differences in schemas, e.g., differences in attributes of two schemas),
- Semantics (related to the differences in intended meaning of terms in specific contexts).

Although this is a useful subdivision, it can still lead to practical misunderstanding situations. Therefore, is useful for better comprehension to approach the different needs for data harmonization through a list of undistinguished causes of heterogeneity in spatial data (Portele *et al.* 2007). Heterogeneity in the case of spatial (geographic, atmospheric, geological) data is for example caused by differences in:

- data format and data collection procedures
- spatial reference system
- data/conceptual model: structure and constraints – metadata model
- nomenclature, classification, taxonomy, terminology/vocabulary, thesaurus, ontology
- scale, degree/amount of detail, extent (spatial, thematic, temporal)
- portrayal (legend/classification, style)
- processing functions: their parameters and formulas/algorithms

3 THE HUMBOLDT WAY

As shown, data harmonisation is an enormously complex task, as there are highly different factors that need to be addressed. At the same time, the organizational environment for geoinformation is also complex, with hundreds, maybe thousands of organizations, either legally mandated or private commercial, with a well-founded interest in harmonised geoinformation. These complexities necessarily are reflected also upon the structure and methodological approach within the HUMBOLDT project.

Therefore, a two-pronged approach was chosen: on the one hand, 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.

On the other hand, the different stakeholder groups have to act as drivers, to ensure that the technological development fulfils their requirements as good as possible. These groups of interest are mainly addressed via the so-called HUMBOLDT scenarios. These represent different GMES application fields, ranging from border security to urban planning, risk management and the protection of nature reserves (Reitz *et al.* 2006).

Other than these two major channels by which the influences and expectations towards the project are managed, there are several additional measures: the collaboration with other projects, the integration of a Review and Advisory Board, the close cooperation with various INSPIRE teams and finally the establishment of both a developer community and the USER@HUMBOLDT platform.

To transform these influences into workable approaches and actual software products, an iterative specification and implementation methodology was chosen. This process also takes into account the heterogeneity of the HUMBOLDT consortium in terms of the location of the teams, the different languages that are spoken and the different technical backgrounds from which the individuals working on the project come from. Since the establishment of such a process is a relatively complex task and needs to be monitored and adapted during a long-running project such as this one, a whole work package (WP04) is devoted to the development of this process and of the interfaces to internal and external users. The input provided by this work package is then used by the specification work package (WP05) to define a functional and technical specification, consistent with itself and with the requirements imposed by the project environment. An additional element of specification is handled in WP07, the data harmonisation work package. Here, both methodologies and models for data harmonisation are researched and fed back to the other specification and implementation work packages.

The actual development work is done in WP08 and WP09, respectively responsible for the framework and the scenario implementation. Additional work packages (WP06 and WP10) ensure the quality of all deliverables, both specification and implementation.

Between all those organizational units of work, well-defined interfaces exist as well as a plan that foresees a total of five major development iterations, of which each one is again subdivided into smaller iterations that can be tracked and managed in a controllable way. Figure 2 gives a view on this iterative process and the various influences that are handled within the HUMBOLDT project.

It is the principal aim of this project to support and advance the process of definition and implementation of the ESDI, contributing to the harmonisation process and introducing scenario results of the framework application, with respect to GMES topics and activities, fulfilling INSPIRE principles and rules and following the requirements of environmental agencies, policy-makers and other users.

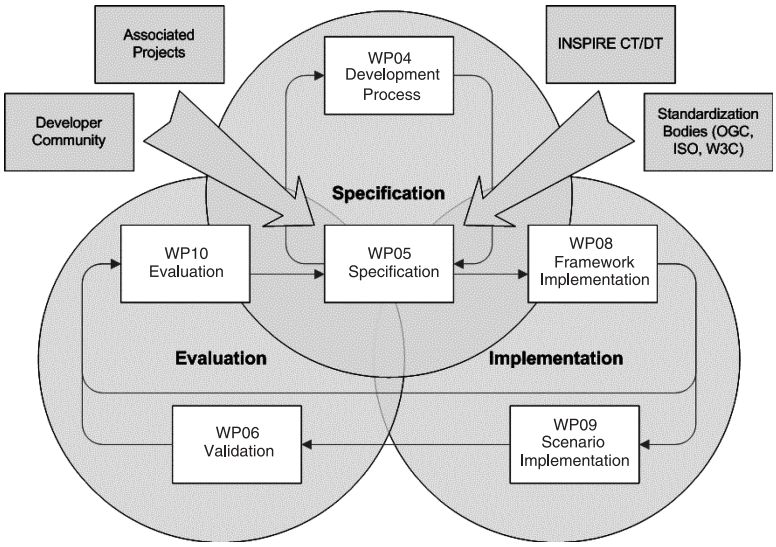

Figure 2. Schema of the iterative process structure regarding HUMBOLDT framework and scenarios.

The role of HUMBOLDT in data harmonisation of geoinformation and in long-term sustainability of GMES is therefore quite important. This importance needs to be addressed with the help of the spatial data user community to complement the research and academic institutions, as well as industrial and administrative structures which are partners in the project.

HUMBOLDT partners are therefore uniting in an effort to fruitfully put into practice data harmonisation experiences, software architectural studies and spatial data tools implementations, with the purpose to set a common harmonisation frame to test and further develop, eventually aiming at usefulness and efficiency of ESDI for European environment and space policies.

4 BENEFITS OF HUMBOLDT

The HUMBOLDT framework for the harmonisation of spatial data will enable mapping and conversion between different metadata thesauri and between different informational classes, or ontologies, thus bringing benefits to geoinformation community of users and developers for geographic services across the European Community (Bernard *et al.* 2005).

Through the solutions of the difficulties and nodes listed above, the proposal of the HUMBOLDT framework will affect the implementation of a European Spatial Data Infrastructure (ESDI) and its community of users, ranging from institutional and political users, to enterprises working in the fields of spatial information, to private citizens and groups, to research communities in geo-spatial data issues.

In particular, the resulting harmonisation of spatial data and services, gained with the HUMBOLDT implemented framework, will enable or at least make easier geographic applications that cross (Friis-Christensen *et al.* 2005):

- National borders;
- Application domains;
- Limitations inhered in spatial data availability, from incompatible data formats to semantic gaps related to lacking data and metadata models.

Moreover, HUMBOLDT has the role of putting a part of the INSPIRE principles and implementation specifications into practice to make sure that guidelines given by INSPIRE are actually implementable.

Once implemented, then, the HUMBOLDT framework will provide interoperability between spatial information systems, thus bringing huge benefits for user's community:

- Enable access to geospatial services not available or not usable at this very moment, using current technological solutions;
- Creation of new information through the access to additional data and services, affecting the decision-making process and making it more comprehensive;
- Enhancement and facilitation of data and services access and distribution, thus making ESDI creation efforts more attractive for commercial and industrial partners.

The choice and production of Open Source solutions during the project will result in benefits and support for spatial data users, for the implementation work is made more manageable for two reasons: the integration of existing knowledge and tools, characterized by open access to the public (especially for non-experts), and the linked reduction of costs; reduction achieved not only for the implementation phase, carried on by HUMBOLDT consortium, but above all for the further development of the framework after the project conclusion, thanks to a open source developer's community.

Finally, HUMBOLDT will be connected to standardization bodies (OGC, ISO, CEN); during the implementation of the framework, international standards and technical specifications will be taken into account not only regarding the exploitation of existing standards, but above all dealing with HUMBOLDT contribution to current standards amendment and future standards definition.

5 SUMMARY & OUTLOOK

This paper has described the background and motivations, the aims and challenges and especially the benefits of the HUMBOLDT project, giving a rationale for the HUMBOLDT framework capabilities and functionality in the context of the implementation of an ESDI.

The major aim of HUMBODT is the implementation of efficient, cost-effective, reliable, generic, interoperable and sustainable solutions for the issue of spatial data harmonisation and integration of geographic services in the framework of an ESDI. This objective is to be reached by putting INSPIRE principles into practice, applying

international standards and using as core reference the users' requirements and needs, finally establishing a community of users and developers, composed by research partners and public institutions and private companies, which ensures the endurance sustainability of the HUMBOLDT framework well after the formal end of the project in 2010.

The methodology of the HUMBOLDT development is based on a dual approach, comprising both a technological and an application side, and on an iterative process of implementation, during which the solutions found are tested and validated with the cooperation of an application momentum, composed of scenarios which cover topics of utmost importance in GMES.

The HUMBOLDT project shows challenges both to geosciences research, covering topics in data harmonisation at a continental scale, and to economic and political management of such a large and heterogeneously composed consortium of partners. Nonetheless, the more relevant the challenges to face, the better the benefits which will surge from their solutions: benefits for specialised and non specialised users of spatial data, for policy-makers, planners and managers, for European citizens and their organisations, at a level which varies from local to regional to European are to be delivered through HUMBOLDT, during the years to come.

REFERENCES

Annoni, A., Smits, P.C. 2003. Main Problems in Building European Environmental Spatial Data, *International Journal of Remote Sensing* 24 (20) 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 Infrastructurs (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 29 15–31.

De Vries, M., Giger, C., Loidold, M. 2007. HUMBOLDT – State of the Art in Data Harmonisation and Data Management. *Deliverable A3.5D1 of HUMBOLDT project* (March 2007).

Friis-Christensen, A., Schade, S., Peedell, S. 2005. Approaches to Solve Schema Heterogeneity at the European Level. *11th EC-GI & GIS Workshop*, Alghero (Italy).

Marchetti, P.G. 2007. Advancing earth observations missions and geospatial interoperability within the Heterogeneous Missions Accessibility project. Geophysical Research Abstracts, European Geosciences Union 9 04799.

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* 19 (1) 85–95.

Toth, K., Nunes de Lima, V. 2005. Data Quality and Scale in the Context of European Spatial Data Harmonisation. *11th EC-GI & GIS Workshop*, Alghero (Italy).