ISSUES OF MULTILINGUALITY IN CREATING A EUROPEAN SDI – THE PERSPECTIVE FOR SPATIAL DATA INTEROPERABILITY

Joanna Nowak, Javier Nogueras-Iso, Stephen Peedell
Institute for Environment and Sustainability, Joint Research Centre
I-21020 Ispra, IT
{joanna.nowak | javier.nogueras | stephen.peedell}@jrc.it

ABSTRACT

In July 2004 the INSPIRE Proposal for a Directive was adopted by the Commission, aiming to establish an Infrastructure for Spatial Information in Europe. The first thematic application domain that INSPIRE will address is environment and in such a context an efficient implementation and monitoring of environmental policies requires interoperable spatial information across national borders and streamlined access and use of this information by all concerned stakeholders. Shared understanding between independently developed heterogeneous geospatial databases applications is not a new problem. In the European context, however, the development of a Spatial Data Infrastructure must embark the issue of multilinguality, which increases the complexity of facilitating the interoperability of such data. This paper identifies the different aspects where multilinguality may influence the development of a European Spatial Data Infrastructure.

KEYWORDS: European Spatial Data Infrastructure (ESDI), Infrastructure for Spatial Information in Europe (INSPIRE), mulitilinguality, interoperability.

INTRODUCTION

In July 2004 the INSPIRE Proposal for a Directive was adopted by the Commission, aiming to establish an INfrastructure for Spatial InfoRmation in Europe [EC, 2004]. The provisions of the proposed Directive cover metadata, spatial data sets and services, network services, agreements on sharing, access and use, coordination and monitoring mechanisms processes and procedures. The ESDI Action of the European Commission Joint Research Centre (JRC) has the task to technically co-ordinate the INSPIRE initiative and the various steps working towards the realization of a European Spatial Data Infrastructure. Since the first thematic application domain that INSPIRE will address is environment, an efficient implementation and monitoring of environmental policies requires interoperable spatial information across national borders, its streamlined access and use by all concerned stakeholders.

Figure 1 shows the classical Geospatial Access Resource Paradigm that is presented in the Global Spatial Data Infrastructure Cookbook [Nebert, 2004]. Here, this paradigm has been intentionally framed in the European context. In this context, the issues of multilinguality represent an additional difficulty for interoperability in every phase of this Geospatial Resource Access Paradigm: the resource discovery, the resource evaluation, and the resource access.

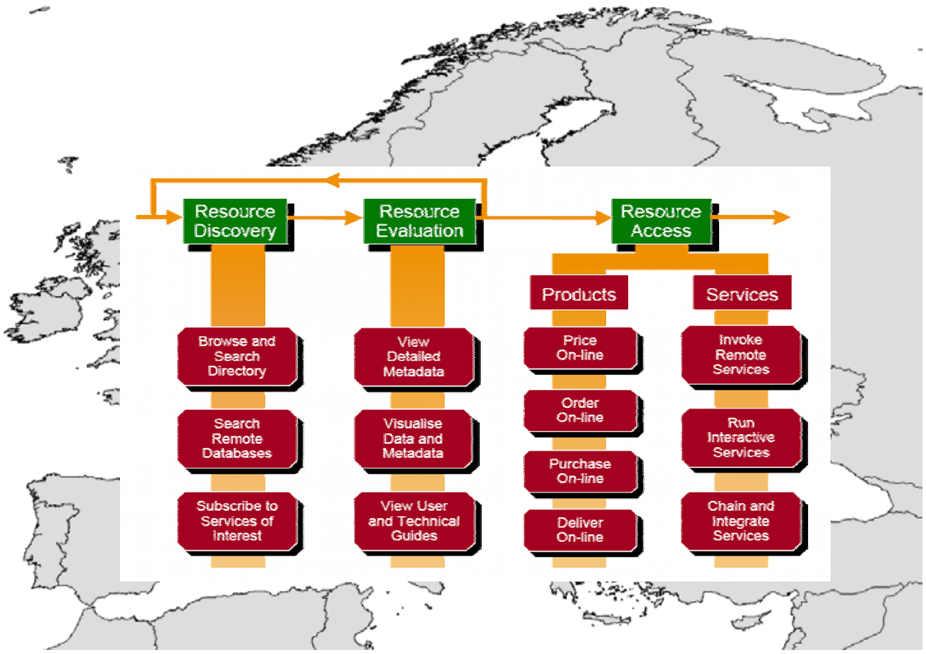

Figure 1: Geospatial Resource Access Paradigm, modified from [Nebert, 2004]

As concerns multilingual resource discovery, several issues must be taken into account. For instance, the graphical user interface of geoportals should be internationalized. Furthermore, these geoportals these geoportals must help the users to establish their query constraints in multiple languages.. Regarding this point, the INSPIRE directive proposal already recommends in Chapter III, Art 13 (2c) the key attributes and the corresponding multilingual thesauri for terminology and taxonomy harmonization [EC, 2004]. Additionally, a strategy for cross-language information retrieval must be developed. As stated in [Bernard et al., 2005], multilingual aspects of metadata should be considered within the European context. Member states are not expected to provide translation for each metadata record they produce and thus a European SDI catalogue must tackle the problem of finding resources independently of the language used for metadata and data creation. Therefore, cross-language information retrieval strategies could consider both the automatic translation of queries to all possible languages, the automatic translation of metadata documents to all possible languages, or the indexing document and queries in some common and language independent representation.

Regarding multilingual resource evaluation, a European Spatial Data Infrastructure (ESDI) should comprise the viewing of detailed metadata in another language, and provide enough means to visualize the data appropriately. Although the visualization of data seems language independent, SDI developers must worry about the internationalization of legends and the display of internationalized attribute information if necessary. For instance, the BALANCE project [Ostländer et al., 2005] uses external XML files to provide the translations of Web Mapping Services (WMS) capabilities documents, which are used by the client for the translation of WMS data layers. Moreover, in the phase of resource evaluation, other multilingual and multinational issues must be taken into account, e.g. the selection of the correct Spatial Reference System, or the appropriate symbology according to the cultural traditions of each country.

And finally, multilingual resource access and further processing concerns the general problem of spatial data interoperability. The exchange of scattered information is related to the heterogeneity of its sources. In a distributed environment which consists of a large number of independently developed geospatial databases that refer to different world views one can find different representations, different schemas, terminologies and hence different semantics [Bishr, 1997]. It can be deduced, the multilinguality context of a European SDI influences strongly these heterogeneities, especially terminology and semantics.

The objective of this work is to analyse in further detail the heterogeneities that originate the multilingual spatial data interoperability in this third phase of the geospatial resource access paradigm, i.e. when resource access and further processing of different sources are required. This paper defines the problem of interoperability, the implications in the geographic information context, and the implications of the multilinguality as its additional dimension.

MULTILINGUAL SPATIAL DATA INTEROPERABILITY

The term interoperability has been developed within the geographic information context as "the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units" [ISO, 2005]. Formerly, the Institute of Electrical and Electronics Engineers (IEEE) defined interoperability as the "ability of two or more components to exchange information and to use the information that has been exchanged" [IEEE, 1990].

The notion of interoperability involves a communication process between two agents (i.e. machines or human beings) that use their own terminology to abstract their subjective world perceptions [Brodeur and Bédard, 2001]. Basing on an analogy with human communication, one can observe (or experience) that the different world views and the use of different terms may lead to serious problems when the intuitive meaning of the elements from the terminology is not clear and the understanding of the information provider differs from that of the user.

Given above, one can notice that the definition of interoperability is data-driven. Within the more specific context of Geographic Information, our intention is to enable spatial data sharing that is any spatially referenced and time referenced information about geographic features.. Consequently, spatial data interoperability can be defined as the ability to access, share and manipulate spatial data stored in heterogeneous distributed repositories. The restriction to the spatial data does not necessarily decrease the complexity of the problem of interoperability, since it usually requires the involvement of different scientific disciplines such as communication or cognition, not to mention geosciences or computer science.

Spatial data interoperability poses many difficult challenges at multiple levels. Often the organizational, operational, and technical interoperability is distinguished. In this paper we focus on another typical fine grained distinction that is syntactic, schematic (called also structural), and semantic interoperability. The latter division originates with different heterogeneities that arise during information sharing. According to [Bishr, 1997], the syntactic heterogeneity is the difference in software platforms, the geometric representational, and topologic relationships of spatial objects; the schematic heterogeneity is the difference in database models or schemas; while the semantic heterogeneity is the difference in context information. In line with the more recent approaches [Shekhar, 2004], the syntactic interoperability specifies common message formats (e.g.

tags and marking) to interchange spatial data, patterns and relationships; and the structural interoperability provides means for specifying semantic schemas for sharing; while semantic interoperability involves an agreement about content descriptions of spatial data, patterns and relationships.

Europe is a patchwork of large numbers of countries and regions characterised by their own culture and languages. After the last EU enlargement (May, 2004), there are twenty European official languages in the Member States. Moreover, some countries are multilingual (i.e. Spain, Belgium, and Luxemburg). This multicultural reality sometimes causes problems since culture is a sphere least capable of being subject to common policies and harmonisation.

Terminology is an inseparable issue from the huge differentiation within world due to the cultural assumptions we all make, as well as historical, geographical and social differences. Terminology, a structured set of concepts and terms of a specific subject field in a specific language, is also used to store the spatial information in databases. As a result, the independently developed geospatial databases that refer to different, subjective world views and use different representations, different schemas, terminologies, and languages cause the misunderstandings and communications problems in the process of sharing data. And that is why it is a great challenge for a European SDI to manage the multilingual spatial data interoperability, i.e. provide the ability to access, share and manipulate of spatial data stored in heterogeneous distributed repositories in different languages.

The interoperability of spatial data primarily requires identification and understanding of the characteristics of heterogeneity. Therefore, the more detailed consideration in this paper is towards an understanding of both the causes and examples of structural (schema), syntactic (format) and semantic (meaning) heterogeneity in the multilingual environment.

Syntactic and Structural interoperability

Syntactic and structural interoperability support the handling, exchange and combining of data properly, with regard to formats, encodings, properties, values, data types and so forth. Regardless the standardisation efforts of the organizations like the Open Geospatial Consortium (OGC), software developers do not necessarily include the OGC specifications and data creators use different schema languages. Consequently, different syntax is used to represent data (different communication protocols, encodings, data formats). Although multilingualism seems to hardly influence the interoperability on the syntactic level, the subjective, culture-driven world views and various conceptualizations of spatial objects imply some differences in their computer representation (raster vs. vector, polygon vs. polyline).

Structural heterogeneity means that different information systems store their data in different structures [Wache et al., 2001]. Different communities or expert groups are likely to describe topographic features in different manners. Various categories or instances definitions, as well as geometric descriptions, are the cause of the differences in hierarchies, classes, or attributes, or geometry. For example, the schematic heterogeneities arise when a particular feature is classified under different object classes in different databases, or when an object in one database is considered an attribute in another. Such variations in database models or schemas limit spatial data interoperability. Even the common differences in the geometric representation and topologic relationships of spatial objects can hinder the process of data sharing, e.g., when corresponding objects are represented as polygon in one data set while in the other one they are represented as point or linear objects, or when equivalent objects are characterised by different geometric resolution or referenced to different geodetic datum. Different ranges of allowable values are also examples of schema heterogeneity.

In the multilingual distributed environment some problems of structural interoperability become particularly important. One of them is called vagueness of the instant definition and it occurs- when non-taxonomic relationships (and thus subjective) are used to define a feature. For instance, in a Czech spatial database one can define a feature called. 'reka' for the description of 'river' and a feature called 'potok' for the description of 'smaller rivers or streams'. However, the subjective definition of big rivers and small rivers may be completely different in another language and cultural context.

Another typical conflict is the difference in the level of features generalization. This occurs when one class that is defined individually in a database can be included in other database as part of a broader class, which is semantically similar but a more general one. For instance, in one database one can find a fine grained feature for 'roadside crosses and little chapels' ('krzyże i kapliczki przydrożne' in Polish database) that can be included in other databases within the more general feature 'cultural heritage monuments'. Once again, this fine grained distinction is based on the cultural history of a country, i.e. the importance of a specific type of monument may influence the creation of an individual feature that is not usually considered separate in other countries. This also happens between a database that defines 3 different features for 3 real objects (e.g., 'rzeka', 'kanal', and 'rów odpływowy', Polish terms for 'river', 'channel', and 'drain' respectively) and another database that defines only one feature called river with a type attribute to differentiate several real objects (e.g. the attribute type can have 3 different values: river/channel/drain). Another similar example could be the existence of a feature called 'roads' versus the unique existence of the broader feature 'inland transport networks'.

Additionally, other conflicts may occur in the differences found in the class attributes, i.e. when two semantically similar classes are characterized by different lists of properties. For instance, a feature called 'city' and defined for a Spanish database may include the attributes province code and region code (a region aggregates several provinces). However, a feature defined in other countries without a higher subdivision in regions includes only a unique code for an equivalent concept to provinces.

It is worth mentioning that some problems can be considered as schematic/structural problem by one group of researchers while the others can view them as semantic problems, for example the different levels of granularity [Ceci et al., 2004; Kuhn, 2004]. Often the granularity levels reflect the different world views and, in consequence, influence the hierarchies (taxonomies) differ with respect to their depths.

Semantic interoperability

Imagine that you are faced with a data set that you can display however you are not able to understand the information you know exists within. This is a possible scenario when schema and format heterogeneities are solved.

Semantic heterogeneity is the difference in context information. It occurs when the same real world entity has different meanings in different databases. Due to intended meaning of terms in a special context or application, the data is often of limited use in another context [Bernard et al., 2003]. Human cognition of the world, often based on human perception and social agreements, is subjective. The same rule concerns the process of object conceptualization, including surface and the relations among its man-made and natural features. Moreover, geographic categories are much more likely to show cultural differences in category definitions than are the manipulable objects of table-top space [Smith and Mark, 1998; Mark and Turk 2003a,b]. Thus, social agreements on semantics based on naming of natural geographic features (e.g. mountain, lake, bay) as well as man made features (path, road, channel) are essential part of GI.

Semantic interoperability must provide the shared understanding of spatial information to be exchanged and resolve any ambiguity. Towards multilingual (and thus multicultural) information sharing, the following instances of semantic heterogeneity are to work out: naming and cognition conflicts.

Figure 2: An example of naming heterogeneity

Naming conflicts occur when naming schemes of information differ significantly. This phenomenon appears when two (or more) semantically similar classes have different names, in two independent contexts. For instance, the concept of ‘lake’ is expressed as ‘lago’ in Spanish, ‘tó’ in Hungarian, and ‘jezioro’ in Polish (see Figure 2). Moreover, this level of synonyms’ relationships can be enlarged through the second level. By the second level of this linguistic relationship, we mean the sets of synonyms that exist within particular languages, e.g.: ‘flod’, ‘elv’, ‘å’ (Danish terms for ‘river’) or ‘tanque’, ‘lagoa’, and ‘depòsitos de àgua’ (Portuguese terms for ‘pond’).

The opposite case, i.e. the use the same or similar terms, for non identical information, introduces homonyms that represent the conceptual (or cognitive) heterogeneities [Kuhn, 2005]. For instance, when the term 'lake' comprises 'natural lake', 'artificial fish-pond' and 'inundate lake' (equivalences in Polish language: 'jezioro' for 'jezioro', 'staw rybny', and 'zalew'). Another example of many concepts behind a given term represent 'village green' when it is used for the following concepts: 'meadow', 'grassland', 'lea', and 'hayfield'. These instances illustrate the relationships between hypernym (broader term) and hyponyms (more specialized term) as well. It is worth mentioning there that GI discovery based on keywords (string-matching) is not able to take into account such linguistics-based relationships or any other information ambiguity [Lutz, 2005; Kuhn, 2005].

CONCLUSIONS

This work has presented how mutilinguality issues affect the development of a European SDI. Specially, the focus has been put on the multilingual spatial data interoperability that is required when different resources must be further processed. It has been shown that most of the issues of spatial data interoperability have their analogues in every language used to store and maintain databases. Thus, the amount of topics dealing with the spatial data harmonization and interoperability within the multilingual environment is growing exponentially with respect to the number of languages used. The specific of spatial data interoperability is the consequence of cultural differences in categorizations that are more likely to be found for geographic entities [Smith and Mark, 1998]. We should properly use technology to overcome these differences and preserve different world views (vide globalization threads) and domain specific knowledge.

Open Geospatial Consortium International Organisation for Standardisation as well as CORBA UML have addressed and facilitated many issues of syntactic and structural interoperability related to spatial data [Nogueras-Iso et al., 2005, Bernard et al., 2003]. Also many works concentrate on providing support for overcoming differences in meaning during the discovery and retrieval of geospatial information in the open and distributed environment [Bernard et al., 2003; Klien et al., 2005; Lutz, 2005; Wache, 2001]. These developments and works can also be applied to multilingual data resources and applications. Yet, even if we provide the methods and tools that solve the problem of finding (discovery) and then accessing (retrieval) of the suitable information, there is still a huge amount of work toward the full understanding of the content of the retrieved data.

And finally, bearing in mind that apart from information, architecture and standards, and policies, people are the main components of Spatial Data Infrastructure, we should put the special attention on them. And the priority should be given to the users. Users of a distributed database system need not be aware of the location and functioning of the parts of the database with which they work. All they are looking forward to is to find, get and process the data they need, according to their own perception and abstraction of the topographic reality [Brodeur et al., 2005]. How to reach users needs and expectations? How long it will take? How much it will cost? These are some open questions that must be still answered in this field. And a proper definition of user abstractions manners, their requirements and needs, would help to solve the multilingual spatial data interoperability. Given that, one can admit that there is still a fair amount of work towards creating multilingual semantic interoperability.

BIBLIOGRAPHY

Bernard, L., Einspanier, U., Haubrock, S., Hübner, S., Kuhn, W., Lessing, R., Lutz, M. and Visser, U. (2003). Ontologies for Intelligent Search and Semantic Translation in Spatial Data Infrastructures, *Photogrammetrie - Fernerkundung - Geoinformation 2003*, vol. 6, pp. 451-462.

Bernard, L., Annoni, A., Kanellopoulos, I., Millot, M., Nogueras-Iso, J., Nowak, J., Toth, K., 2005. What Technology does INSPIRE need? - Development and Research Requirements. *Proc. of the 8th AGILE conference on geographic information science.*

Bishr, Y.A. (1997). Semantic Aspects of Interoperable GIS. *ITC Publication* no 56, Ph.D thesis, Wageningen Agricultural University and International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede, the Netherlands. ISBN 90 6164 1411.

Brodeur, J. and Bédard, Y. (2001). Geosemantic Proximity, a Component of Spatial Data Interoperability. *The International Workshop on "Semantics of Enterprise Integration", ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA 2001)*, Tampa Bay, Florida, 14th October 2001.

Brodeur, J., Bedard, Y. and Moulin, B. (2005). A geosemantic proximity-based prototype for the interoperability of geospatial data. *Computers, Environment and Urban Systems,* in press.

Ceci, M., Appice, A. and Malerba, D. (2004). Spatial Associative Classification at Different Levels of Granularity: A Probabilistic Approach. *Knowledge Discovery in Databases: PKDD 2004, Lecture Notes in Artificial Intelligence*, vol. 3202, pp. 99-111.

EC (2004). Proposal for a DIRECTIVE of the European Parliament and the Council of the European Union of 23 July 2004 on establishing an infrastructure for spatial information in the Community (INSPIRE), available at: http://inspire.jrc.it/sdic_call/EN.pdf

George, D. (2005). Understanding Structural and Semantic Heterogeneity in theContext of Database Schema Integration. *Journal of the Department of Computing, University of Central Lancashire UCLAN*, Preston UK, no. 4.

IEEE (1990). *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*. Institute of Electrical and Electronics Engineers, New York, NY.

ISO (2005). Geographic information – Services, ISO19119:2005. International Organization for Standardization (ISO).

Klien, E., Lutz, M. and Kuhn, W. (2005). Ontology-Based Discovery of Geographic Information Services - An Application in Disaster Management. *Computers, Environment and Urban Systems (CEUS)* (in press).

Kuhn, W. (2003). Semantic reference systems. *International Journal of Geographical Information Science*, vol. 17, pp. 405–409.

Kuhn, W. (2004). Semantics of What?. *Report of the NCGIA Specialist Meeting on Spatial Webs, Santa Barbara, December 2-4 2004*, National Center for Geographic Information and Analysis, University of California.

Kuhn, W. (2005): Geospatial Semantics: Why, of What, and How?. *Journal on Data Semantics III*, 2005. In Press

Lutz, M. (2005). Overcoming Differences of Meaning in Spatial Data Infrastructures – Achievements and Challenges. Position Paper for *Workshop on Cross-learning between Spatial Data Infrastructures and Information Infrastructures, Enschede, the Netherlands*.

Mark, D.M. and Turk, A.G. (2003a). Ethnophysiography. *Workshop on Fundamental Issues in Spatial and Geographic Ontologies, 23 September 2003 (in conjunction with COSIT03)*, http://www.geog.buffalo.edu/ncgia/ethnophysiography/PreCOSIT-Ethnophysiography.pdf

Mark, D.M. and Turk, A.G. (2003b). Landscape Categories in Yindjibarndi: Ontology, Environment, and Language. in Spatial Information Theory - Foundations of Geographic Information Science. *COSIT 2003, Kartause Ittingen, Switzerland*. 2003: Springer, Lecture Notes in Computer Science 2825: pp. 31-49.

Montello, D. and Freundschuh, S. (2005). *Cognition of Geographic Information*, in *A research agenda for geographic information science*, R. McMaster and E. Usery, Editors. 2005, CRC Press: pp. 61-91.

Nebert, D. (ed) (2004). *Developing Spatial Data Infrastructures: The SDI Cookbook v.2.0*. Global Spatial Data Infrastructure (GSDI), http://www.gsdi.org.

Nogueras-Iso, J., Zarazaga-Soria, F. J. and Muro-Medrano, P.R. (2005). *Geographic Information Metadata for Spatial Data Infrastructures - Resources, Interoperability and Information Retrieval*. Springer Verlag, ISBN: 3-540-24464-6.

Ostländer, N., Tegtmeyer, S. and Foerster, T. (2005): Developing an SDI for time-variant and multi-lingual information dissemination and data distribution. *Proc. of 11th EC GI&GIS Workshop, ESDI: Setting the Framework, 29th June - 1st July 2005, Alghero, Italy*.

Shekhar, S. (2004). Spatial Data Mining and Geo-spatial Interoperability. *Report of the NCGIA Specialist Meeting on Spatial Webs, Santa Barbara, December 2-4 2004*, National Center for Geographic Information and Analysis, University of California.

Smith, B. and Mark, D. M. (1998). Ontology and Geographic Kinds. In T. K. Poiker and N. Chrisman (eds.), *Proc. 8th International Symposium on Spatial Data Handling (SDH'98), Vancouver: International Geographical Union, 1998*, pp. 308-320.

Sowa, J. (2000). Ontology, metadata, and semiotics. In *Ganter, B. and Mineau, G., editors, Conceptual Structures: Logical, Linguistic, and Computational Issues*, pages 55–81, Berlin. Springer-Verlag.

van Harmelen, F., Patel-Schneider, P.F. and Horrocks, I., (eds) (2001). DAML+OIL ontology markup language. DAML. Org, Reference description, http://www.daml.org/2001/03/reference.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H. and Hübner, S. (2001). Ontology-based Integration of Information - A Survey of Existing Approaches. In: *Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle*, WA, 2001, Vol. pp. 108-117.

(Draft) Proceedings of the 11th EC-GI&GIS Workshop: ESDI: Setting the Framework. 2005, p. 47-48

Institute for Environment and Sustainability

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Abstracts

ESDI: Setting the Framework

11th EC-GI&GIS Workshop

Alghero, Sardinia
29 June-1 July 2005

www.ec-gis.org/Workshops/11ecgis

INSPIRE · Infrastructure for Spatial Information in Europe

REGIONE AUTONOMA DELLA SARDEGNA

Table of Contents

11th EC GI & GIS Workshop, ESDI: Setting the Framework