Data consistency and multiple-representation in the European Spatial Data Infrastructure

Katalin. Tóth

European Commission, DG Joint Research Centre

katalin.toth @jrc.it

Abstract
In Spatial Data Infrastructures the “multiple” principle is widely used in information management. This paper presents what are the objectives of the INSPIRE Directive (2007/2/EC) in fields of multiple representation and data consistency, how the legal and technical framework to be established contributes to coherent environmental information management.

Keywords: INSPIRE, data specifications, multiple representation, data consistency

1 Introduction

With the world wide spread of the Geographic Information Systems the paradigm of single purpose data production has changed. The possibility of re-using and combining data and information from various sources created the need to constitute a framework where people can communicate, access and use data without specific efforts of humans and machines. Such a framework, called Spatial Data Infrastructure (SDI) encapsulates policies, institutional and legal arrangements, technologies, data that enables sharing and effective usage of geographic information [Craglia at al 2003]. SDI can be established at global, supra national, national, regional, cross-border, and local levels[1]. In ideal case these levels are connected accommodating the relevant components of the others.

Being its most expensive part data has crucial role in information infrastructures. The cost of data collection accounts for around 60-80% of the total costs of setting up a system.[2] Good metadata, powerful network services, and clear rules for data sharing pave the road for re-use, but the usefulness of information depends on how much data is comparable with the actual needs in terms of semantics, schema, format, and data matching.

“Multiple” is a key concept of SDI. The entities of the real world are described from multiple points of view, with multiple levels of details; the same data is addressed to multiple users in multiple products tailored according to their requirements. However “multiple” the production and the application process is, coherence of information is needed that on its turn requires consistency of data. The objective of SDIs is to supply data through interoperable services based, when applicable, on harmonised data.

In INSPIRE the “multiple” concept of SDI has been recognised; proposing “multiple representation” and “data consistency” as harmonisation elements for the data component.

The recent paper summarises the latest achievements of INSPIRE related to the use of multiple-representation and the related data consistency issues, synthesising the results of the data specification process[3] and the key findings of the workshop[4] organised by the Spatial Data Infrastructures Unit of the Joint Research Centre.

1 There is no commonly accepted sub-division of SDIs. Regional SDI is equally used at supra and sub-national levels. The author proposes to use the term „supra-national” when the SDI is supported by appropriate legal agreement. Cross border SDI can be established by neighbouring regions of two or several nations.

2 Source: PANEL-GI Compendium, EUR 19360en, 2000

3 Preparatoy phase 2004-2006 focused on the conceptual framework of specification development

2 The INSPIRE directive

2.1 Background

The INSPIRE Directive (2007/2/EC of 14 March 2007) lays down the general rules aimed at the establishment of the **IN**frastructure for **SP**atial **I**nfo**R**mation in the **E**uropean Community for the purposes of the Community environmental policies and the policies or activities which may have an impact on the environment. As a piece of European legislation a Directive sets out the objectives to be achieved, while does not regulate how to do so. Having entered in force on 15 May 2007 the Directive has to be transposed in the national legislations of the Member Stated within 2 years, and parallel to the legislative process terms on technical and legal aspects of implementation have to be agreed within 5 years. These arrangements will be defined in the Implementing Rules for INSPIRE, drafted with wide involvement[5] of the stakeholders and subject to the regulatory procedure[6].

Member states are free to decide how to comply with the Implementing Rules. In case of interoperability of spatial datasets and services they may choose to use on the fly services to implement the necessary conversions and transformations or can store the results off-line to be delivered upon a specific respect. Methods and procedures are out of the scope of implementing rules; however they can be made available for the community as best practice examples.

The building blocks of INSPIRE are metadata, interoperability of spatial data sets and services, network services, data sharing and implementation monitoring. It is expected that INSPIRE will have a wider effect in Europe; however the use-cases behind this infrastructure will stem from environmental applications. ISNPIRE shall be built on the SDIs established and operated by the Member States of the European Union; it does not require collection of new data.

2.2 Data consistency requirements in the Directive

Based on the assumption that multiple representations of the real world entities (both from multiple points of view and multiple scales/resolutions) are extensively used the INSPIRE Directive sets explicit requirements on data consistency:

Article 8

1. In the case of spatial data sets corresponding to one or more of the themes listed in Annex I or II, the implementing rules […] shall meet the conditions laid down in paragraphs 2, 3 and 4.
(…)
3. The implementing rules shall be designed to ensure consistency as between items of information which refer to the same location or between items of information which refer to the same object represented at different scales.
4. The implementing rules shall be designed to ensure that information derived from different spatial data sets is comparable as regards the aspects referred to in Article 7(4[7]) […].

Article 10

2. In order to ensure that spatial data relating to a spatial feature the location of which spans the frontier between two Member States are coherent, Member States shall, where

4 INSPIRE Multiple-Representation and Data Consistency Workshop. 24 participants, 15 invited experts from 8 countries (Czech Republic, France, Finland, Germany, Hungary, The Netherlands, Switzerland, and UK) representing National Mapping and Cadastral Agencies, industry and academia.

5 Stakeholders may porpose well-acepted specifications as candidate ones for INSPIRE, may submit reference material, delegate members to the Drafting Teams and the Thematic Working Groups and participate in the review of Draft Implementing Rules.

6 In accordance with Article 5 (metadata and monitoring)and 5a (data specifications, network services and data sharing) of Council Decision 1999/468/EC of 28 June 1999

7 Definition and classification of spatial objects and the way of geo-referencing

appropriate, decide by mutual consent on the depiction and position of such common features.

The above paragraphs require that the INSPIRE Implementing Rules address the issues of multiple representation and data consistency. The data consistency can be one of the aspects of “spatial data validity” to be documented in the metadata (Article 2, paragraph 2(c)).

The same real world object depending on the intended use can be modelled from different (thematic) views and within each view with different level of details. Multiple-representation, both intentional and unintentional, is an everyday practice in geographic information technology. The change of paradigm that has replaced map production with information management forced data producers to adopt new production lines, replace single data collection by multiple data uses with heavy use of generalising information [Stoter, 2005]. Never the less parallel data collections continue existing, inevitably creating a risk that consistency between different representations is not always guaranteed. Consequently INSPIRE has to address these issues when technical provisions for implementation are defined.

3 Data consistency in INSPIRE

3.1 Conceptual framework

One of the indicators of SDI implementation is the degree of data harmonisation. Shared information schema (elements C, D. E in fig.1) and adequate operational components (K, L, N, S) open up the way to harmonisation, never the less specific modelling (R) and data consistency measures (Q) are still required.

In INSPIRE 20 elements are proposed for data harmonisation (fig.1). These elements are formalised in the INSPIRE Generic Conceptual Model (GCM). The GCM is the basis for developing the application schemas for each of the 34 data themes listed in the three annexes of the Directive. This application schema, together with other specification elements of ISO 19131:2007 (Data product specifications) serve as basis for the data specifications that will be legally binding for the Member States as Implementing Rules of INSPIRE.

(A) INSPIRE Principles	(B) Terminology	(C) Reference model
(D) Rules for application Schemas and feature catalogues	(E) Spatial and temporal aspects	(F) Multi-lingual text and cultural adaptibility
(G) Coordinate refe-rencing and units model	(H) Object referencing modelling	(I) Data translation model/guidelines
(J) Portrayal model	(K) Identifier Management	(L) Registers and registries
(M) Metadata	(N) Maintenance	(O) Quality
(P) Data Transfer	(Q) Consistency between data	(R) Multiple representations
(S) Data capturing	(T) Conformance	

Fig.1: Data harmonisation elements in INSPIRE

For information integration common coordinate reference system is needed (component G). Object referencing (H) can support a seamless micro to macro views, information aggregation and dis-aggregation contributing to data integrity and consistency. Sound data integrity should be coupled with visual coherence at all resolutions (J), as people need to view information coming from different sources across different resolutions in a consistent way.

In order to provide a harmonised view to concepts of spatial object types, attribute types, association types and coded values a common feature concept dictionary as specified in ISO 19126 (Feature concept dictionaries and registers) shall be maintained for INSPIRE in an ISO 19135 conformant register. The role of a common feature concept dictionary for all INSPIRE data specifications is in particular to support the harmonisation effort and to identify conflicts between the specifications of the spatial object type in the different themes.

No concept shall be modelled as part of a INSPIRE application schema if it is competing with a concept already established as part of the Generic Conceptual Model. Similarly, all concepts which are of general utility and not limited to a theme shall lead to a change proposal for the Generic Conceptual Model and should not be modelled in a INSPIRE application schema. Only additional concepts of the thematic area must be modelled as part of the application schema for a given data theme. However application schemas may have dependencies between each other. These dependencies stem from cross thematic consistency requirements that should be addressed during the application modelling.

3.2 Application schema development

EN ISO 19109:2005 proposes a straightforward way for application schema development. Depending on the actual interest of the (potential) users the Universe of discourse is modelled in meta model with feature types that are later represented by the appropriate elements of the spatial schema. The feature types can be either genuine concepts to be documented in the feature catalogue or can be equally taken from an existing one. From SDI point of view this second has to be followed whenever possible, which leads overall consistency in the infrastructure. The classification of the real world objects should be done without overlaps and gaps; feature and attribute definitions must be unambiguous.

The ISO 19109:2005EN standard, being high level and generic does not address one of the key points of correct modelling: the selection rules. The semantic resolution, or “scale of reasoning” [Ruas, 2002] defines the granularity or the level of details (LoD) of the model. The same object may be essential in one application, while worthless in another. Similarly the same phenomenon may appear at object or aggregation level, with simpler or more complete attribute set.

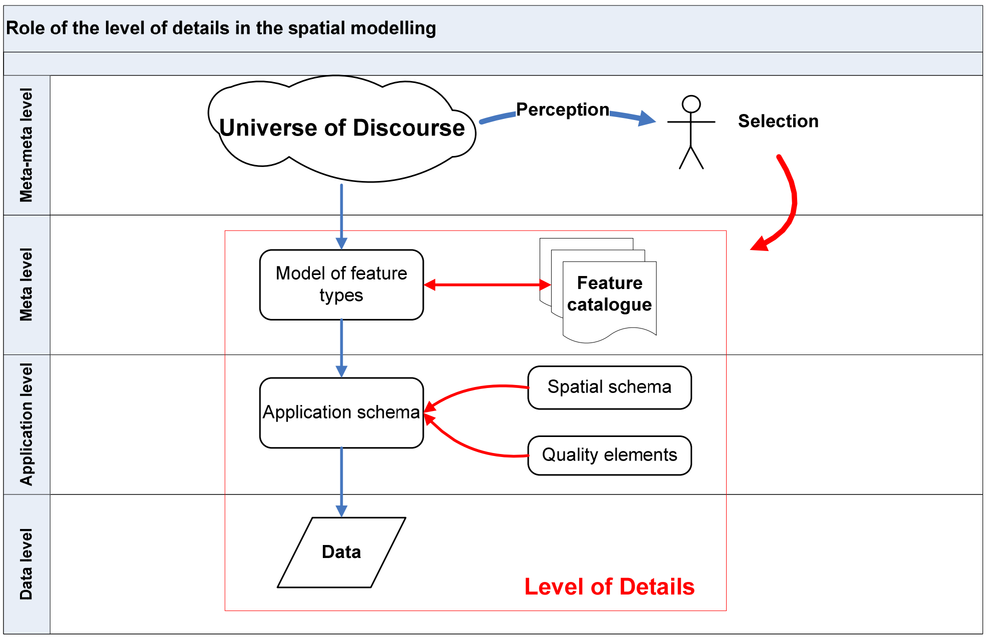

Fig.2: Modelling process as proposed in ISO 19109:2005 EN completed with the role of the level of details

The level of details defines how to describe spatial objects[8] in terms of the spatial schema and the "acceptable" quality that in their turn define the applicable data collection methods as well. Therefore the whole modelling - system implementation – data collection cycle should be governed by the targeted LoD. This dependency is shown in fig.2 proposed by [Tóth and de Lima, 2005]

It is not difficult to see that there is no theoretical limit for increasing the semantic resolution. Never the less each model has its own limit. This limit usually is connected to spatial resolution level, however spatial resolution is not the only classifier in this process (e.g. landmarks – like churches, single trees, etc. - are not deselected, reclassified or aggregated).

The actual level of details for each data theme will be defined based on the analysis of the user requirements. Spatial Data Interest Communities are invited to participate not only from the part of data providers, but also from users. Based on selected Community policies it can be predicted that the required scales/Levels of Detail (LoDs) may span from large (noise maps at street level according to the Directive 2002/49/EC) to small (spatial representation of the river basins according to Directive 2000/60/EC). Recent SDI studies show on one hand that a selected data theme may be available at different LoD from country to county; on the other hand a data theme can be covered by diverse datasets even within a country that should be considered when data are to be harmonised.

An application schema as defined in [2] shall contain a complete and precise description of the semantic content of its spatial object types. The spatial object types and their properties specified in an application schema shall be drawn from the common feature concept dictionary. Whenever this is not possible it is reasonable to start from the existing situation, which makes easier the implementation and reduces the costs. The European Feature Catalogue produced by the EuroSpec project together with the inventory of features/attributes of topographic domain may give a valuable input to the application specifications. It is important to underline that the classification of the real world objects should be done without overlaps and gaps; feature and attribute definitions must be unambiguous.

Data consistency cannot be separated from correctness. Users/data providers may wish to know how well the actual implementation of a dataset describes the Universe of Discourse. Data modelling has crucial role in defining the selections rules, and the corresponding appropriate geometric and attribute (thematic) accuracies, object classification, geometric representation and topology.

Correctness of information is of utmost importance for the users. It depends not only on the quality of data per se stored in the systems, but also on how the data describe reality. Geographic reality cannot be measured exhaustively because it is nearly impossible to obtain measurements for every point across the entire landscape Therefore a fundamental discrepancy exists between geographic data and the reality that they are intended to represent. This discrepancy or uncertainty is propagated through and may be further amplified by data management [Tóth and De Lima 2005]. Good starting point that means appropriate modelling has determining role in the quality of geographic information.

For a single data theme depending on the LoD several application schemas can be developed in INSPIRE, however this must be justified by appropriate use-cases. Basic principle should be to use as few LoD as possible, which simplifies not only the data modelling, but also organisational and management aspects of maintenance.

Consistency requirements change with the scale, thus generalisation gives a reasonable framework for their modelling. Taxonomic and partonomic relations that are expressed in metric, topological and longitudinal terms give quantitative measures for consistency too. The experience accumulated in cartographic science in this field gives valuable contribution for building SDIs [6].

8 In the INSPIRE Directive the term "spatial object" is used insteaad of "feature".

3.3 Consistency within the data theme

As it is stated in the previous part, whenever possible, only one application schema is foreseen for a single theme. In this case the consistency within the data theme can be approached through the compliance to the specifications of the theme. Before checking the consistency between different data sets, each data set has to be verified against the corresponding INSPIRE application schema, in particular compliance with the same set of constraints. Object Constraints Language (OCL) may be useful for checking logical consistency. Other specification elements like data quality that are expressed by measures, allow the users to compare how much an implementation of a dataset adheres to the specifications.

Consistency levels should be diversified for reference and thematic data setting stringent requirements for the firsts and being more relax for the second. Consistency should be analysed in different aspects (e.g. positional consistency and semantic consistency) and the requirements must be diversified depending on the concrete data.

If there is a need for specifying not one, but two, or even more application schemas for the same data theme, it is reasonable that the spatial object types of different levels follow generalisation-specification hierarchies, possibly formalised in multiple-representation data models.

Preserving links between different representations contributes to update propagation, thus to data consistency. In incremental update a small change at higher level of detail can cause global changes at smaller level of details. As the update propagation is not straightforward a small change in the larger scale may have a global impact, especially in aggregation [6].

3.4 Cross theme consistency

People need to view information coming from different sources and at different resolutions in a consistent way. In environmental applications there is a need to combine different thematic information. Obvious consistency constrains occur for example in case of digital elevation models and hydrography, cadastral parcels and administrative units, addresses and buildings, etc. An initial set of related themes was identified during the INSPIRE theme definition and scoping, which will be completed based on the user requirement survey before starting the data specification process.

Themes with strong inter-relations have to be grouped and specified together. It must be explicitly stated that two datasets are comparable in terms of consistency and the specific rules should be developed and formalised using constraints. However, consistency must be enforced only in justified cases. Requirements how to build the schema should be proved by valid use-cases of Pan-European and cross border scenarios.

The consistency between different themes should be checked only within the same or close levels of detail. From semantic point of view there is a need for an interlinked and agreed cross-domain vocabulary. Correspondences between the databases can be established by various tools of data matching (data mining, ontologies and formal specifications) and transformations (conflation, generalisation, matching geometries). Never the less mapping data in a common schema usually requires manual work and adapting objects/attributes from additional sources

To improve the consistency between data of different legacy one or both data sets may need to be slightly changed. These data transformations will require mutual approval from data providers and also a technical organisation to match the data together either before sending them to the user (an on-demand matching service) or once, before integrating them into INSPIRE.

3.5 National boundaries

Last, but not least in INSPIRE data consistency problems occur along the state boundaries when combining data provided by neighbouring states. These problems span from diverging classification of spatial entities, through different level of details, legal aspect (different border line registered in the interested countries) to data accuracy problems. The latter can be caused by the measurement methods resulting different positional accuracy, but also by the errors in the transformation parameters when converting coordinates from one projection system to another. The custodians of such spatial datasets, as required in the Directive, will decide by mutual consent on the depiction and position of the common spatial objects or they will agree on a general method for edge-matching or other automatic means to maintain data consistency [2].

State boundaries preferably must have a strong legal basis formalised in international treaties and a common datum – ETRS89. In case of dispute, as last resort, suitable data matching algorithms can be used. This fixed frame of state boundaries may serve for anchoring other datasets from both sides. Intersections between the state boundaries and the linear elements of the transport network or hydrography can be marked and used for segmentation before generalisation - establishing a sufficiently rigid frame for data matching.

4 Conclusion

Data coming from multiple users describing the real world from multiple points of view at multiple levels of details has to be presented to variouis groups of users. In spite of the complexity of the integration process the information supplied to the users must be coherent. The INSPIRE Directive sets up the framework of a spatial data infrastructure for environmental policy making in the European Union. As part of this data harmonisation is foreseen as an applicable way to achieve interoperability of spatial datasets and services.

Data harmonisation in INSPIRE means not only harmonising the application schemas for each theme listed in the annexes of the directive, but also dealing with aspects of multiple representation and data consistency. Data consistency is required in different context; within a dataset, between different levels of details, between interdependent data coming from various domains, and finally, in case of spatial objects that span over the national boundaries. The decision about the implementation aspects of INSPIRE is in the competence of the Member states. However methods, algorithms and procedures can be shared with the wider community as best practice examples.

The INSPIRE Directive and the related Implementing Rules create not only a technical framework for coherent environmental information management but also the legal basis for implementation in each Member State of the European Union.

References

Balley, S. et al (2004) Modelling geographic data with multiple representations. International Journal of Geographical Information Science Vol. 18, pp 327 – 352

Benchikha, F. Boufaida, M. (2005) Roles, Views and Viewpoints in Object –Oriented Databases http://swt.cs.tu-berlin.de/~stephan/VAR05/papers/BenchikhaBoufaida.pdf

Craglia M. at all (2003) Geographic Information in the Wider Europe. http://www.ec-gis.org/ginie/doc/ginie_book.pdf

Friis-Cristensen A. (2003) Issues in the conceptual Modelling of Geographic Data. Phd. thesis, Department of Computer Science, Aalborg University

Jakobsson, A. (2005) Reference Data Sets and Feature Types in Europe. http://www.eurogeographics.org/eng/documents/ref_data_ver101_part_A.pdf and http://www.eurogeographics.org/eng/documents/ref_data_ver101_part_B.pdf

Kavouras, M. Kokla, M. (2006) Geo-ontology Integration: Identifying Issues, Dimensions, and Developing Guidelines. Presentation at Joint ISPRS WG II/3 – WG II/6 Workshop on Multiple Representation and Interoperability of Spatial Data, Hannover

Stoter, J. E. (2005a) Generalisation: The Gap Between Research and Practice. 8th ICA Workshop on Generalisation and Multiple Representation, La Coruña 7-8th July. http://ica.ign.fr/

Stoter, J. E. (2005b) Generalisation within NMAsin the 21st Century. International Cartographic Conference La Coruña 9-16 July

Ruas, A. (2002) Spatial analysis and Agent principles to automate generalisation process http://geopro.cic.ipn.mx/Pruas.pdf

Tóth, K., Nunes de Lima, V. (2005) Data Quality and Scale in Context of European Spatial Data Harmonisation (11th EC-GIS workshop, Alghero http://www.ec-gis.org/Workshops/11ec-gis/papers.cfm)

[1] 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) http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML

[2] INSPIRE draft Generic Conceptual Model http://www.ec-gis.org/inspire/reports/ImplementingRules/inspireDataspecD2_5v2.0.pdf

[3] EN ISO 19109:2005 Geographic Information - Rules for Application Schemas

[4] ISO 19131:2007 Geographic Information - Data Product Specifications

[5] EuroSpec project web-site, Eurogeographics http://www.eurogeographics.org/eng/03_projects_eurospec.asp

[6] INSPIRE multiple-representation and data consistency workshop http://sdi.jrc.it/ws/multiple_rep