

DATA QUALITY AND SCALE IN CONTEXT OF EUROPEAN SPATIAL DATA HARMONISATION

Katalin Tóth, Vanda Nunes de Lima
European Commission Joint Research Centre, Ispra, Italy

ABSTRACT

The proposal for the INSPIRE Directive sets out the legislative framework for implementing the European Spatial Data Infrastructure as based on the national ones. One of the elements of the future infrastructure is the harmonised spatial data and spatial data services. Keeping in mind the heterogeneity of the potential input, the compatibility of the conceptualisation and the quality of the datasets are pivotal. Scale and data quality are not only metadata, but also instruments in this process. Moreover, the distributed client server environment sets new challenges both to the metadata documentation and the quality assurance. After a historical overview and problem statement the paper proposes some pragmatic solutions.

KEYWORDS: Data harmonisation, scale, quality, metadata

INTRODUCTION

The proposal for a Directive of the European Parliament and the Council establishing an Infrastructure for Spatial Information in Europe (INSPIRE) sets out a legislative framework for the implementing of the components of a European Spatial Data Infrastructure. This infrastructure includes metadata, harmonised spatial datasets and interoperable services. According to the directive the implementing rules “shall be designed to ensure that it is possible for spatial data sets to be combined, or for services to interact, in such a way that the result is a coherent combination of spatial data sets or services that represents added value, without requiring specific efforts on the part of a human operator or a machine.”

The European Spatial Data Infrastructure must be based on the infrastructures of the member states. As a consequence of the historical heritage mapping in Europe has developed according to the political reality – state boundaries separated not only nations, but knowledge and practices. As recent GIS solutions were built up step by step on traditional mapping, we still have to face the diversity of projection systems, the data and business models and the different rules for access and copyright.. However the questions of global world require global answers based on temporal and spatial extent. The integration of multidisciplinary knowledge further complicates the question of diversity.

The INSPIRE lays down general rules for the establishment of an infrastructure for spatial information in Europe to support environmental policies or policies that effect the environment. One of the elements of this infrastructure is harmonised data. Given the huge number and the cross-discipline character of the data themes, sustainable cost-effective methods are necessary for the implementation.

REENGINEERING VERSUS. HARMONISATION

The straight way of achieving harmonised spatial data is to agree on a clear semantics and the common conceptual model. Semantics is necessary to clarify the understanding of the terms, while conceptual model specifies the way of the abstraction of the universe of discourse: the objects to be represented, the way how these objects should be described in terms of properties and the spatial schema, and finally some constraints related to the features itself or related to their relationships.

Naturally, the agreement on the conceptual model depends on the intended use of the product; the requirements of the user define the level of details, the applicable data collection methods and the conceptualisation itself. This direct scenario can be followed when products are developed from the scratch, or when the interested parties agree on reengineering their data. Implementing this approach requires considerable resources and time, not speaking about the difficulties in negotiating the specifications themselves.

The other approach relies more on the technology, considering that a good deal of work can be solved by using the tools schema translation. Never the less harmonisation is still necessary. The questions of semantics still remain the subject of accord. The INSPIRE proposal foresees not only the definition and classification of spatial objects, but also their key attributes; spatial representation and relationships, the way of geo-referencing; how information on the temporal dimension is exchanged and updated. In the heterogeneous European environment strict standardisation process is hardly feasible. What probably can be done is to deliver specifications that serve as a basis for comparison in the harmonisation process. Therefore harmonisation means compliance to the agreed specifications.

THE ROLE OF SCALE IN CONCEPTUALISATION

The INSPIRE foresees the harmonisation of 31 strongly application driven datasets of the Member States. Even for a single theme there is a variety of sources that differ in conceptualisation, spatial representation and quality. The driving force of the conceptualisation is the intended use, which defines both the level of details and the terms for the abstraction. It is well known from traditional cartography that the same object can be represented as a line or a surface; a symbol or a sign in scale. Similarly small objects present on large scale maps are not shown individually on a small scale one. This process is the cartographic generalisation, which consist of selection and geometric simplification.

The term of scale cannot be interpreted directly in digital cartography. In case of raster data the resolution meaningfully describes the level of details. In case of vector data scale can be used only as a metaphor to show the analogy of the conceptualisation. Never the less it facilitates the discussion and the comparison between different levels of details. We should not forget either, that many of the digital products were derived from analogue maps. Why not to use then the scale in a pragmatic way in digital mapping – so called “equivalent scale”?

Creating the application schema for digital spatial data principally follows the same way. But what happens if we want to harmonise data? First thing is to agree on the reference model. In case of INSPIRE the reference model should reflect the requirements set up by the users’ group of the Spatial Data Interest Communities (SDICs). As it is shown in fig.1 they have to define the necessary scale (level of details) of the data. This is the base to select the way of representation of the features from the spatial schema, and naturally the corresponding data quality. Until now there is no accorded cross reference between the scale and the corresponding quality of the data. The European data harmonisation process perhaps cannot overlook this problem.

In order to preserve the inclusiveness of the European data harmonisation process it seems to be reasonable to agree on several mandated levels of details with the corresponding reference models. It is necessary to define the functions of mapping between the levels of representation in such way that the topological relations between the objects are preserved and cause minimum changes in the geometry.

Once the reference specifications are in place, data custodians have to define which of their existing datasets may satisfy a certain specification requirement. This implies checking the consistency in the content, the conformity with spatial representation and accuracy requirements, and the completeness of the attributes.

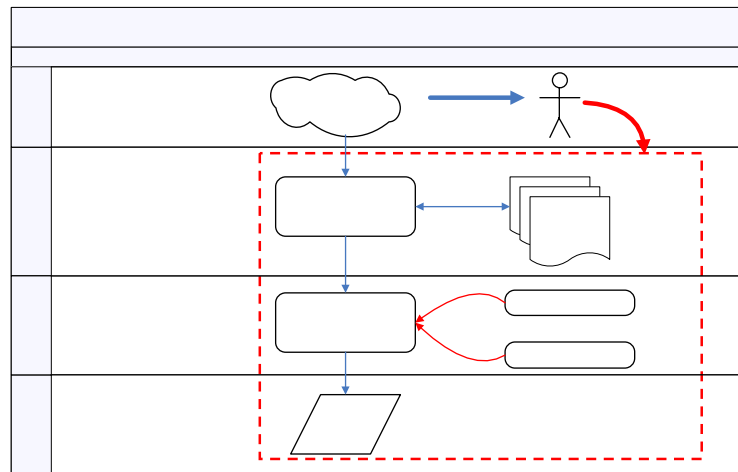


Figure 1: The role of scale in application development

There are different views whether datasets need to be transformed to a common scale. No doubt, it is much more convenient for the users if they do not need to address data heterogeneity during their analysis and decision making. The multiple representations of features allow the users to make the choice to deal with homogeneous data or with the partially increased accuracy. At that stage, it is important also to consider that data transformation is always associated to error(s), in most of the cases unknown their results and how they propagate.

QUALITY OF GEOGRAPHIC INFORMATION

In traditional mapping the objects that depict the real world are often defined by regulations, giving the obligatory content according to the intended use, the selected scale, and the desired quality of the end product, which fixes in its turn the applicable technology too. The legally mandated levels of accuracy are achieved through the allowed measuring methods. Data collection, editing and portrayal form a unique technological process with high human involvement and continuous quality control. The integrity of the end product is maintained by the

(practically) non-alterable base, by the paper itself. Metadata like scale, measuring method, producer, time of publishing, etc. is documented on the information carrier, therefore data and metadata integration the data authenticity is ultimately solved and data authenticity is guaranteed by the publisher.

The similar well-established system for digital data collection has lagged behind the technology. According to McGlamery (2001) when information is transferred to digital format the issues of integrity and authenticity are overlooked. In many cases the analogue map compilation procedures have been adopted in the application schemas for different products, paying more attention to the spatial than to the quality schema. As soon as the data is taken out of the initial context, the integrity of the quality and the intended use is not guaranteed any more. The readily available functions like zooming in a dataset, resampling, or the relatively easy way of attaching attribute from other datasets may tempt the user to freely merge datasets of divers conceptualisation.

Data quality is a pillar in any GIS implementation and application as reliable data are indispensable to allow the user obtaining meaningful results. Data transfer, sharing and integration are common practices by many users. In GIS applications, data from different sources, with different levels of accuracy and with differences in quality can be combined. All data sources and spatial data entry methods present errors into the information created and used for display and analysis. The type, severity and implications of these errors inherent in a GIS database determine the quality of spatial data. These errors should be recognised, documented and properly dealt with.

Identifying and assessing data errors are not the only factors which determine data quality. Data quality is relevant in all the processes involved with conceptualisation, developing, utilizing and maintaining a spatial database. These include data collection, data input, positional and attribute accuracy, data storage, data manipulation, data conversion and quality control procedures. The data quality in the Geographic Information Systems is often understood only as metadata, or an a posteriori report about the purpose, use, lineage, completeness, logical consistency, precision, positional, temporal and thematic accuracy of the data. These reports are usually stored separately from the data. This risks on the one hand that they are not maintained according to the data updates, while on the other hand, in case data come from different sources, they may be inconvenient for the user to deal with.

To ensure that existing digital data be appropriately used, the data producer must provide documentation about the “history” of spatial data. In addition to spatial data documentation, data developers and users should document and implement data quality measurements, which allow judgment to be made about spatial data.

Spatial data is frequently relied upon as factual data. Good data quality measures and documentation may eliminate liability law suits against data developers and users. Data producers must be aware of the implications involved with carelessly or non-documented data. On the other hand, the data user must also be responsible for understanding the limitations of that spatial data, to coherently apply the well known “fitness for use” (Veregin 1989) criterion.

The fitness for use varies with the user and the specific user’s need and its perception is fully dependant on the scale, accuracy, and extent of the data set, as well as the quality of other data sets to be integrated. Therefore the harmonisation in the field of spatial data quality is concerning the standard implementation of data quality measurements and the documentation about the spatial data set, which should include data sources, data input techniques, positional accuracy, attribute classification and definitions and quality control procedures used to validate the spatial data.

According to ISO two components of data quality are identified. Data quality overview elements providing informative non-quantitative information and data quality elements providing quantitative quality information that reports how well a data set meets the criteria set forth in its

product specification. Data quality elements include the quality components of completeness, logical consistency, positional accuracy, precision, temporal accuracy, thematic accuracy and allow for the creation of additional user defined components. Each component is comprised of several aspects called data quality sub-elements. Data quality information for each sub-element is reported in several parts, including a data quality scope, data quality measure, data quality evaluation procedure, data quality results, value domain and date. According to ISO, the metadata schema given in 19115 is the mandatory method for reporting data quality information.

Spatial data relate information about three aspects of a geographic feature: typology (the type of geographic feature), location, and spatial dependence. Because such attributes change over time, geographic data are very complex and difficult to manage. Geographic reality often cannot be measured exhaustively because it is nearly impossible to obtain measurements for every point across an entire landscape. Accurate measurements are also difficult to obtain because of continuous (slow or rapid) variation of the landscape over time and because of the limitations of instruments, financial budgets, and human capacity. Thus, when geographic data are developed, they are merely approximations of geographic reality. 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 and analyses in a GIS environment.

The basic GIS schemes (Couclelis 1992) for representing geographic data are not dynamic but record only a static, invariable view of the world. They do not depict complex objects that consist of interacting parts, nor do they display variation at many levels of detail over space and over time. Thus, uncertainty must be recognized as a basic element in all GIS results. Uncertainty analysis assesses the discrepancy between geographic data in GIS, and the geographic reality that the data are intended to represent. The current state of GIS technology in dealing with uncertainty falls short of the goals described by Goodchild (1993, p. 98): (1) each object in a GIS database would carry information describing its accuracy; (2) every operation or process within a GIS would track and report error; and (3) accuracy measures would be a standard feature of every product generated by a GIS.

THE WAY FORWARD

In spite that data may stem from various sources for a single application, data and metadata integration is not an everyday practice in the recent GIS meanwhile the users should know how “good” the data is. Naturally, the strict solution is when quality information is attached to each entity. On one hand data models can open towards the metadata integration. Depending on where metadata reside, e.g. dataset, theme or feature level (Langaas 1995) they can be integrated especially in the object oriented data model at various levels. When a simple feature is taken out from a dataset using some methods the data quality attributes can be inherited. This is highly desirable in the distributed client-server environment.

On the other hand, in the most of the cases not single features, but whole data layers or datasets of similar origin and lineage are to be integrated, thus the quality can be referred to them as a whole. In case of the INSPIRE, where data harmonisation seems to be application driven according to reference models, the users can be sufficiently informed about the data quality if conformity to the specification of a certain representation level is referred. Therefore a conformity checking procedure needs to be elaborated and made available to all the data supplier parties. The advantage of this approach is that larger parts of the data are characterised and there is a higher probability to have a complete set of metadata including not only the quantitative data quality measures, but also the data quality overview elements like the purpose of the creation of the datasets, the measuring methods and the lineage (history of the dataset).

Naturally the users do not want to waste too much time on reading metadata for each element, they need a simple and quick way for communication. Quality visualisation offers an effective “on the spot” judgement about the “fitness for use” and data uncertainties.

A possible scenario for the data harmonisation as based on compliance checking is shown in fig.2.

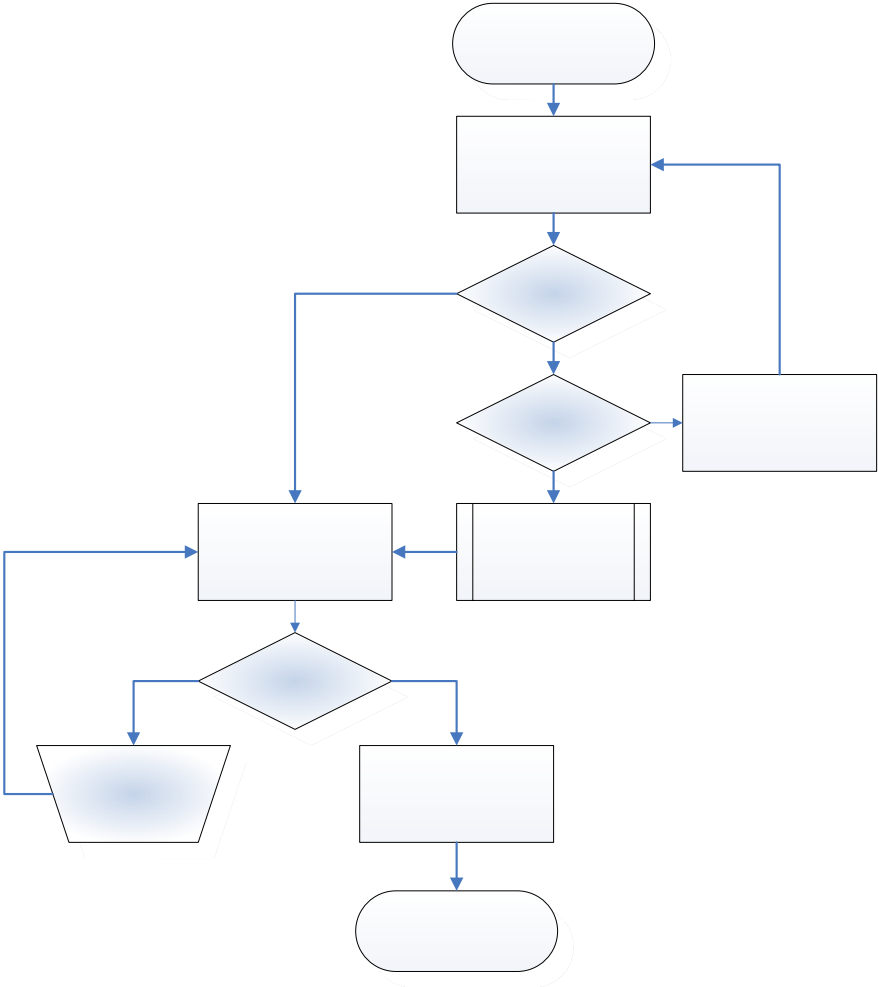


Figure 2: A scenario for the spatial data harmonisation process

CONCLUSION

As INSPIRE foresees the European Spatial Data Infrastructure as built on the existing national infrastructures of the Member States, the quality concept is pivotal not only in the documentation (metadata), but also in the data harmonisation process itself. The distributed data management and the web mapping and feature services require arrangements for the quality assurance of spatial data. The level of details or scale should be treated as the driving force of the conceptualisation, which on its turn defines other quality parameters as well.

A series of interlinked reference models may guarantee the inclusiveness process, enabling each Member State to join the harmonisation according to the quality and availability of the data. The interaction between the different Spatial Data Interest Groups (SDICs) may yield the necessary use cases and the appropriate data models.

BIBLIOGRAPHY

- Borges, K at all: Integrity constrains in spatial databases. Jorge Horacio Doorn, Laura C. Rivero (eds.): Database Integrity: Challenges and Solutions. Idea Group 2002. pp144-171
- Couclelis, H., 1992. People manipulate objects (but cultivate fields): Beyond the raster-vector debate In: U. Frank, I. Campari and U. Formentini, editors, Theories and Methods of Spatio-Temporal Reasoning in Geographic Space (Lecture Notes in Computer Science Vol. 639). Berlin-Heidelberg:Springer-Verlag, pp. 65-77
- Goodchild, M. F., 1993. Data models and data quality: Problems and prospects. In: M. F. Goodchild, B. O. Parks, and L. T. Steyaert, editors, Environmental Modeling with GIS. Oxford University Press: New York, pp. 94-103.
- Goodchild, M. and Gopal, S. (eds.) 1989, Accuracy of Spatial Databases. Taylor and Francis, London. pp. 263-276.
- McGlamery, P. 2001, Issues of Authenticity of Spatial Data. INSPEL 35 2, pp 137-144
<http://forge.fh-potsdam.de/~IFLA/INSPEL/01-2glpa.pdf>
- Langaas, S. and Tveite, H., 1995, To Characterise and Measure Completeness of Spatial Data: A Discussion Based on the Digital Chart of the World (DCW)
<http://ilm425.nlh.no/pub/gis/dcw/quality.ps>
- ISO 19109. Geographic information. Rules for application schema
- ISO 19115:2003 Geographic Information - Metadata