Improving geographical datasets usability by interactive schema transformations

Sandrine Balley
Institut Géographique National
Saint Mandé, France
sandrine.balley@ign.fr

SUMMARY

Acquiring geographical data that best fit an application requirements is a complex task. Choices concerning information structure and representation are as crucial as those concerning information content. Users need assistance in understanding what information content and structure they need, which are available and which can be derived through simple pre-processing steps. We aim at proposing a system enabling users to acquire a custom dataset from available datasets and restructuring operations. The user should interact with the system through a graphical interface displaying the conceptual schema of an available dataset and operations that are possible on this schema. This interface allows the user to derive a schema that best fits his needs and the user eventually retrieves a dataset structured after the schema he has specified. This paper introduces the context and objectives, some related works and the first steps of this work.

KEYWORDS: *Fitness-for-use, data schema, schema transformation*

CONTEXT AND OBJECTIVES

Usability of geographical data

Usability is defined by ISO as "the effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments." Applied to geographical data, the concept can be seen as a complex ternary relationship between data and their characteristics, users and their needs, and processes to be applied on data (Josselin, 2003) (Jahn et al., 2003). In other words, a dataset usability depends on specific characteristics of the dataset with regard to the user application: information content and information structure.

Two steps can be distinguished in the access to a usable dataset.

- The first step consists in identifying the dataset the most adapted to the intended application. This can be done through the consultation of metadata. Current metadata servers enable a wider range of users to make a motivated choice amongst datasets. However, this choice is difficult and may require some assistance (Grum et al., 2004). Furthermore, metadata servers cannot provide access to perfectly suitable datasets if such datasets are not available.
- If no optimally usable dataset was identified, the second step consists in transforming one or several existing datasets into a new one suiting the intended application. These transformations may affect the information content or structure. They can be problematic: indeed, selecting the datasets that will be best transformable and transforming them require technical skills and tools that some users don't have.

A survey about access to geographical information, led in 2002 from the web site of the research department of the French National Mapping Agency (IGN), can help us to focus the user group the most concerned by this usability issue. During this experiment, initially aiming at collecting some user vocabulary, users were first asked to describe their professional activity. Then, they had to

formulate a query to an imaginary intelligent system able to display any knowledge concerning data or data processing. We got approximately 30 responses. As shown on figure 1, the more frequently asked questions by "lambda users" (i.e. who use geographical information only for their leisure activities) concern the use of GI (e.g. how to use a map together with a GPS device). The experienced users (i.e. who sometimes use geographical information for their professional activiy) rather ask questions about data processing (e.g. how to calculate a buffer, an itinerary, or an intermediate contour line). Only the expert users (i.e. whose professional activity focuses on geographical information) ask for data different from that they already have, that would enable them to easily perform applications (e.g. to study noise levels, to measure building volumes, or to process advanced data visualization). This range of users is particularly aware of the usability question and is asking for more usable datasets. In a first time, our approach then focuses on expert users.

Kind of question asked the more frenquently	How to use maps or datasets	How to process datasets	How to find a dataset « such that I could… »
User type	Lambda users	Users occasionally using GI in their professional activity	Users focusing on GI in their professional activity

Growing expertise in the field of geographic information

Figure 1: Different user needs

IGN is the main spatial data provider in France. To improve the usability of an IGN's dataset, a first way consists in providing rich metadata about the dataset (its quality, its feature catalogue, its specifications, etc.), so that expert users can actually identify its possible exploitation. Another way consists in performing pre-processing to adapt the dataset to the intended use and to deliver "à la carte" information. This is the approach chosen in this work.

Objective

Our goal is to enable users to interactively specify pre-processing at the conceptual schema level before ordering any data. Some pre-processing are commonly available: changing the data format or the coordinate system. Here, we focus on filtering the information content and modifying its representation. Our application interface should enable users to specify a dataset fitting their need as precisely as possible, starting from existing datasets. Parallel to the interactive specification, inner procedures should automatically derive the specified dataset for them.

To meet this objective we first work on the description of available datasets: it must inform the user of the current and derivable contents. This first requires a rich metadata model including data specifications, which is not mandatory for ISO 19115 compliant metadata. This also requires to enrich the data schema with some knowledge about the typical operations to be possibly applied on each data type. Last, it must be presented in a clear and user-friendly interface.

We also work on designing tools to transform datasets. They are associated to the operations described on the schema and translate schema transformations into data transformations. Some of these transformations require some back processing to ensure the data still fits its specifications. For example, connected road sections may have to be aggregated if all their attribute values have become equal after an attribute suppression.

This paper is structured as follows: next section explains what kind of data transformations we are considering. The third part provides a state of the art concerning both access to geographical datasets and database techniques issues. The last section provides some concrete elements of our approach.

TRANSFORMING INFORMATION CONTENT AND REPRESENTATION

To list required dataset transformations, we consider that a geographical dataset is made of two major kinds of information: explicit and implicit information.

Explicit information

Explicit information is the data described by the conceptual data schema or by the feature catalogue (i.e., it is represented as a class, a relationship, a feature attribute or an attribute value). Some modifications of the conceptual schema, like the removal of a feature attribute, alter the explicit information content (figure 2a). Some other do not change the explicit information content, but only its representation within the data. For example, let us consider a dataset representing buildings as a single class with a "type" attribute whose values can be "habitation", "industrial" and "commercial". An estate professional may want to concentrate on habitation buildings, because he's going to add a price information on them. He does not care about other types and would like to visualize them just as contextual information. Working with separated classes representing each type of building could be considered as more usable. In such a case, the amount of explicit information in the data is unchanged, but it is represented differently (figure 2b). The explicit information can also be altered modified without the user knowledge, by data format changes. For example, to migrate from a GIS software managing inheritance relations to a poorer one automatically duplicates some feature properties but do not change the information content.

Implicit information

Implicit information does not appear in the data schema as a class, a relationship or an attribute but may be derived. It also may be mentioned in the dataset specifications. An example of implicit information is the non represented geometry of an administrative entity composed of smaller entities with an explicit geometry. The most useful transformation on implicit information consists in changing its representation to make it explicit, i.e. to give it a place in the data schema through the adequate operations. We do not consider deriving information through complex spatial analysis tasks, like areas concerned by flooding risks. We focus on information that can be made explicit through simple spatial analysis operation. For example, if our estate agent wants to represent the "ground surface" information as an attribute of the "habitation building" class, this can be done by computing objects coordinates.

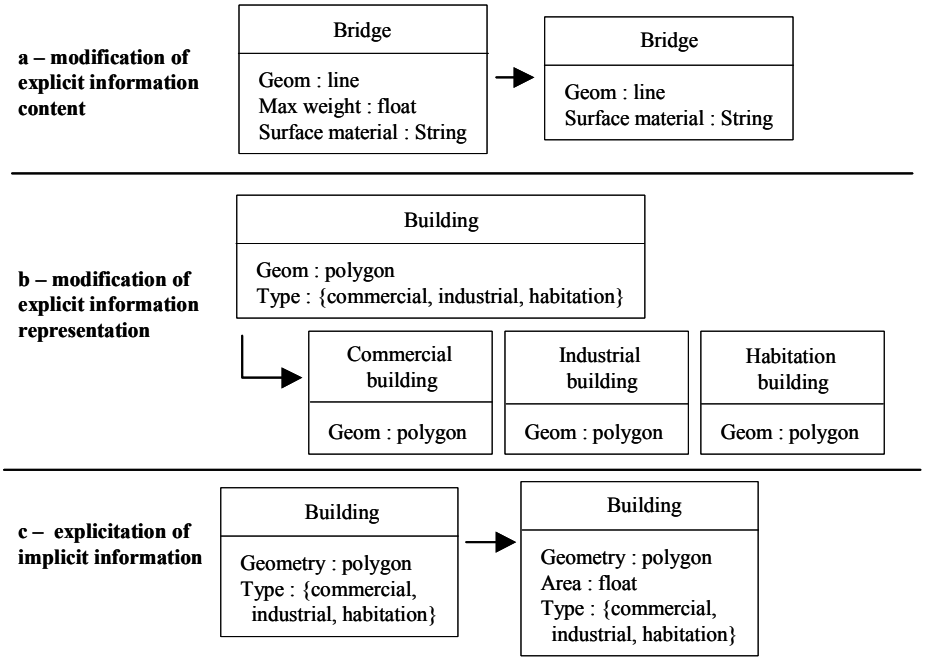

Figure 2: Modifications of information content and representation

Information transformations

Finally, we consider several types of transformation. The first one consists in making explicit some implicit information. The second one focuses on explicit information and consists in transforming the conceptual data schema. The proposed schema transformations are similar to those studied in the domain of classical object-oriented schema evolution (Roantree, 2000): users can rename or remove an object type, a relation type or a property, generalize or specialize an object type, fusion or aggregate several object types, add a relation type or a property. They also can filter the population of an object type by defining predicates on the type's properties. The last transformation type consists of schema evolutions based on spatial and/or implicit information criteria. For example, the explicit "bridge" feature type may be specialized into a new "river bridge" feature type populated by "bridge" instances provided they intersect at least one "river" instance.

RELATED WORK

Systems for data customization

Many works focus on how to measure and express data usability, but only a few focus on customizing datasets to enhance their usability. Some of them adopt a data provider viewpoint: their goal is to modify the dataset before it is sold. For instance, the Research and Innovation center of the Ordnance Survey studies how to let users model their personal dataset by picking some units, possibly more specific than feature types, in an initial dataset (Greenwood et al., 2003).

Another approach, rather chosen by data users, consists in keeping the initial dataset unchanged and building a user view. The concept of spatial view defined by C. Claramunt enables to use the same dataset for different applications. It relies on "atomic spatial views", corresponding to queries, that can be combined (Claramunt et al., 1995). (De Oliveira, 1997) proposes a platform dedicated to

domain expert users wanting to model their own database and application. It includes knowledge concerning common data structures and some design patterns. However, the derivation of the designed dataset from the initial one was not studied. To actually change the data, a GIS software product can be used, but its tools manipulate data selections rather than the complete population of a data schema. Remodeling tools (Seeley, 2003) provided by IBM, SAFE Software or Macromedia can perform powerful modifications on already acquired datasets. A user interface enables to specify these transformations at the logical schema level.

Interfaces based on data schema or metadata

Various applications provide access to data from their schema: most data modelers and diagrammatic querying tools use a rich formalism to represent the conceptual data schema, and mapping rules to bind it with the implementation one.

Many spatial querying languages have been developed for non expert users. They use icons to represent schema elements and some relationships such as geometric or topologic ones (Aufaure-Portier, 1995). Such graphical solutions could be useful in our case to represent the potential implicit information that cannot be displayed in the data schema.

Closer to our topic, some works rely on metadata to evaluate the fitness for use of a dataset. Some of them exploit domain knowledge to automatically detect the best dataset, and even the data manipulation tools adapted to the problem (O'Brien, 2004). Some others, dedicated to expert users, let them explore and derive meaningful information thanks to powerful "metadata mining" tools (Vasseur et al., 2003). Most of the time, these systems rely on one or several ontologies to bridge the gap between the semantics of the user domain and the semantics of the data or metadata.

Schema evolution and views

The major interactions proposed by our system are successive transformations of a conceptual data schema. In the non-spatial database domain, the issue of data schema transformation and evolution has been widely studied. Many evolution operations and their consequences on the schema have been described (Banerjee et al., 1987) (Zicari, 1991), sometimes in the context of view creation (Abiteboul et al., 1991). In our case, these schema transformations will be linked to spatial data transformations, which are also well known and implemented in every spatial DBMS. The implications of schema transformations on data transformation are not always straightforward because of the complexity of spatial data specifications concerning geometry and topology.

OUR APPROACH

System functionalities

Figure 3 proposes in an activity diagram a global process to customize a dataset. We adopt a MVC approach: the user interact with a view of the conceptual data schema. The data schema itself, included in the "metadata" term appearing on figure 3, is transformed as soon as the user has validated the schema view. Data are transformed in consequence. Other metadata transformations affect the "data specifications" part of metadata, which can be annotated if a gap is suspected between them and the data.

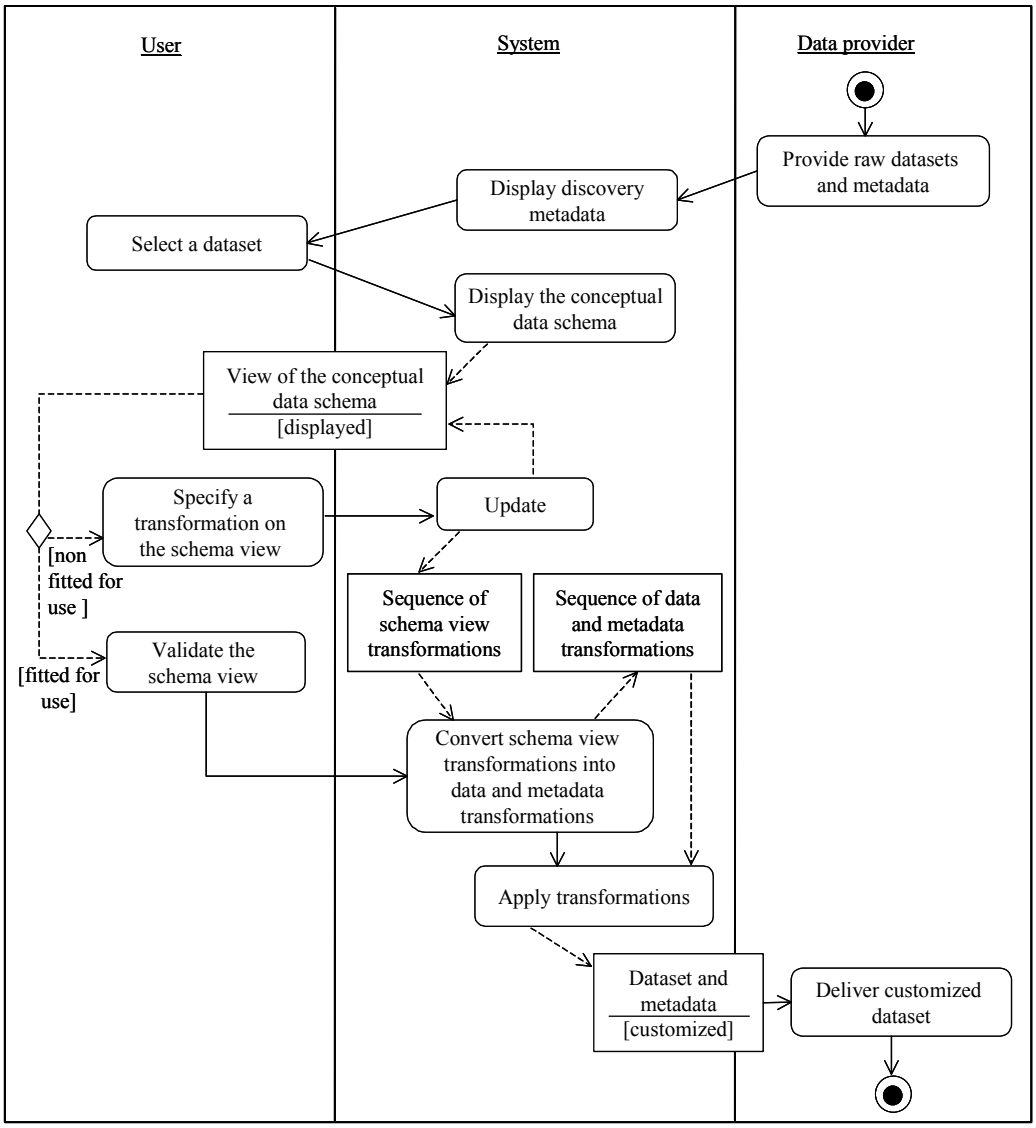

Figure 3: Activity diagram: customization of a dataset.

Resource description model

The resource description model is shown in figure 4 and relies on ISO and OGC models (ISO, 2001) (ISO, 2002) (ISO, 2003) (OpenGIS, 1999). The "Dataset description" part is composed of the conceptual data model (which is proposed to the user), the logical data model (which is the actual manipulation structure of data) and of other metadata. The "Transformation" part is inspired from the geographic tasks model of (Bucher, 2004). Once they are parameterized by the user or by the system, these tasks can be executed relying on simple operations.

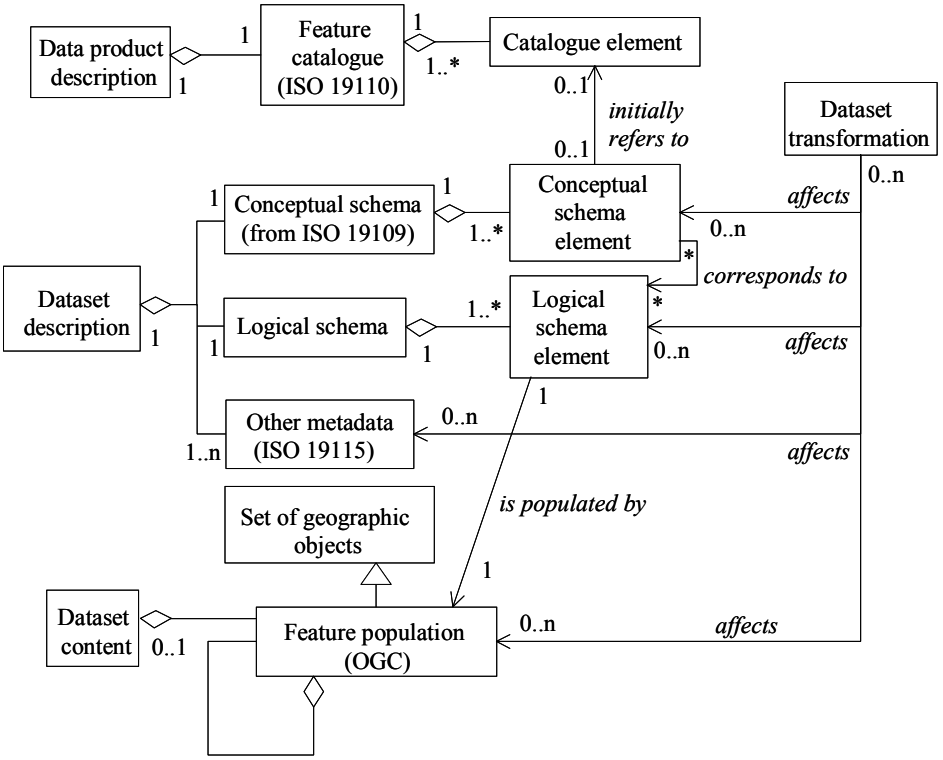

Figure 4: The core resource description model

Implementation

A prototype is being developed on the GeOxygene platform of the COGIT laboratory (Badard et al., 2003). GeOxygene is built on Oracle Spatial DBMS and enables mapping between a relational storage and Java classes representing object types. It was developed in compliance with international standards.

The data description model has been implemented, as well as the automatic graphical display of a data schema from a data dictionary stored in a database. The visualization of the data schema is simplified by the Jgraph Java library. Figure 5 shows a schema display (a) and the contextual user interface appearing when a user clicks on a schema element (b). This interface enables to browse the schema element properties, to select and to execute an applicable transformation amongst the few ones that have been implemented yet.

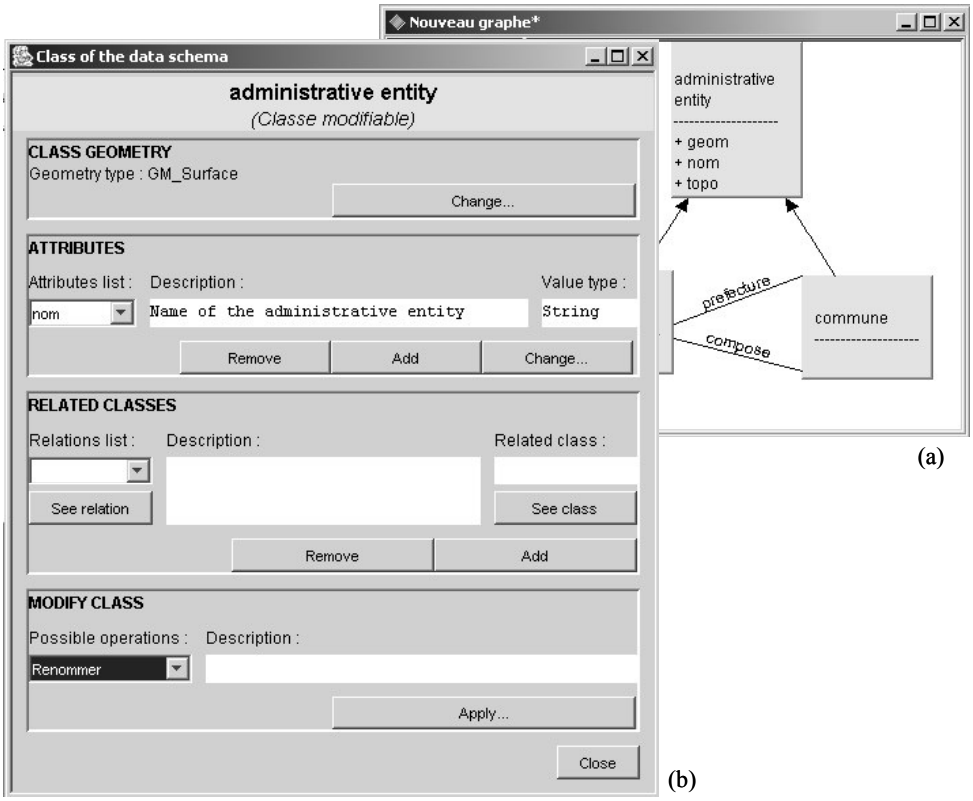

Figure 5: (a) Data schema display using the Jgraph library. Roles cardinalities are formally represented but are not displayed in the current version. *(b)* Object type browsing interface for exploration and transformation.

CONCLUSION AND PERSPECTIVES

This research is focused on the enhancement of datasets fitness for use. We propose an interactive specification of needed dataset and pre-processing including filtering and restructuring of available products.

We intend to do early user tests, focusing first on users IGN, to ensure our prototype is useful for expert users. This will help us to improve the user interface and to validate the list of proposed schema transformations. The prototype is local for the moment, but online solutions should be provided soon

Future work will address the requirements of less experienced users. We expect to define some rules for the validation of the customized dataset, possibly taking into account the user needed application. Some schema transformations are also done automatically when importing a dataset in a GIS software or changing its format. These transformations notably concern the data geometry and topology. Describing the “grammar rules” of some software products to forecast such transformations and to represent their consequences on the final data schema would also be a useful functionality.

BIBLIOGRAPHY

Abiteboul, S. and Bonner, A., 1991. Objects and Views. In Proceedings of the. International Conference on Management of Data. ACM SIGMOD, pp. 238-247.

Aufaure-Portier, M.A., 1995. Definition of a Visual Language for GIS, Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems, T.L.Nyerges et al.(Eds), Kluwer Academic Publishers, (1995) 163-178.

Badard, T. and Braun, A., 2003. Oxygene: an Open Framework for the Development of Geographic Web Services, Proceedings of the 21th International Cartographic Conference ICC'2003, Durban, South Africa.

Banerjee, J., Kim, W., Kim, H.J., and Korth, H.F., 1987. Semantics and implementation of schema evolution in object-oriented databases, ACM SIGMOD International Conference on Management of Data (SIGMOD Record'87). 16, 3, pp.311-312.

Bucher, B., 2004. Integrating structured definitions of processes in geographical metadata. In proceedings of the 11th International Symposium on Spatial Data Handling SDH'2004, pp 629-639.

Claramunt, C. and Mainguenaud, M., 1995, Spatial View: A dynamic and flexible vision of GIS database , Proceedings of the DEXA International Conference and Workshop on Database and Expert System Applications, Revell, N. and Min Tjoa, A. eds, Omnipress, London, UK, pp. 483-493.

De Oliveira J.L., Pires F. and Medeiros C.B, 1997. An environment for modelling and design of geographic applications. GeoInformatica 1, pp. 29-58.

Greenwood, J. and Hart, G., 2003. Sharing feature based geographical information – A data model perspective. In proceedings of the 7th International Conference on GeoComputation, University of Southampton, United Kingdom, 8-10 Septembre 2003

Grum, E. and Vasseur, B., 2004. How to Select the Best Dataset for a Task ? In Proceedings of the third International Symposium on Spatial Data Quality ISSDQ'2004, Frank, A., GeoInfo Series, 28b, pp.197-206.

Jahn, M. and Franck, A., 2003. How to increase usability of spatial data by finding a link between users and data. In proceedings of 7th AGILE Conference on Geographic Information Science, Heraklion.

Josselin, D., 2003, Spatial Data Exploratory Analysis and Usability, Data Science Journal (Spatial Data Usability Section), Voume 2, 26, February 2003

ISO TC211: ISO 19110 Geographical Information – Methodology for feature cataloguing, Final Draft International Standard (2001).

ISO TC21: ISO 19109 Geographic Information - Rules for Application Schema, Draft international standard (2002).

ISO TC211: ISO 19115 Geographic Information - metadata, international standard (2003).

O'Brien, J. and Gahegan, M., 2004. Representing, manipulating and reasoning with geographic semantics within a knowledge framework. In proceedings of the 11th International Symposium on Spatial Data Handling SDH'2004, pp 585-603.

OpenGIS Consortium: The OGC Abstract Specification - Topic 5: Features (1999).

Roantree, M., 2000. Constructing View Schemata Using an Extended Object Definition Language. PhD Thesis. Napier University.

Seeley, R.S.: Data Schemas, 2003. When the Method Becomes the Madness: Issues in Sharing Digital Geospatial Data in a Global Environment and Schema Remodelling Tools as one Viable Solution, In proceedings of the 21st International Cartographic Conference (ICC'2003).

Vasseur B., Devillers R ., Jeansoulin R., 2003. Ontological approach of the fitness for use of geospatial datasets. In proceedings of 6th AGILE Conference on Geographic Information Science, Lyon.

Zicari, R., 1991. Primitives for schema updates in an object-oriented database system: a proposal. Computer standards and interfaces (13),pp. 271-284.