SLOVAK JOURNAL
OF
CIVIL ENGINEERING

2007/2 PAGES 28 – 34 RECEIVED 18. 2. 2007 ACCEPTED 4. 6. 2007

I. IVÁNOVÁ

MODELLING OF THE DATA QUALITY IN THE SPATIAL DOMAIN

Ivana Ivánová, Ing., senior lecturer
Research field: Data quality of spatial databases
Dept. of Theoretical Geodesy
Faculty of Civil Engineering of the Slovak University of Technology in Bratislava
Radlinského 11, 813 68 Bratislava, Slovakia
E-mail: miroslava.matejcekova@stuba.sk

ABSTRACT

The main objective of the article is to introduce quality modeling into the spatial domain with respect to standard definitions. Spatial data have specific characteristics, which enables the assessment of the quality of a spatial dataset from various perspectives – geometric, thematic or temporal, together with distinguishing the producers' and users' point of view on the dataset. The cadastral domain in the Slovak Republic is one of the domains with detailed specification; therefore it appears to be suitable for demonstrating the principles of modeling the quality of spatial data. The detailed research behind this article was a subject of the dissertation – Ivánová, I.: Data quality of spatial datasets (Ivánová, 2006).

KEY WORDS

- *quality,*
- *Universe of Discourse,*
- *ontology,*
- *standardization,*
- *spatial dataset,*
- *quality parameters*

INTRODUCTION

The greatest emphasis in the ”information society” is put on the interchange of information and the sharing of knowledge. The sharing of knowledge may appear in terms of the essential characteristics of the elements of a domain (ontological terms) or in terms of how elements of a domain can be used (problem-solving terms). Spatial information is recognized as a fundamental part of an information infrastructure, which benefits the whole society – in the same way transportation or telecommunication infrastructures do.

A starting point for the creation and utilization of information about the Earth's surface, both its physical-geographic and social-economic components, is a functional spatial model, which is an abstraction. A functional spatial model graduates from reality through 'user-oriented' information into the 'computer-oriented' structure of data storage. The result of this process should be a data model represented by the spatial data itself e.g., in the form of a spatial dataset.

The objective of the research behind the article presented is to design an appropriate method of defining and documenting the quality of spatial data concerning an approach in compliance with international standards related to quality. This objective has arisen from the necessity to efficiently assess and document the quality of spatial data, which is affected by the following facts:

- the reliance of the society on spatial data is growing,
- the widespread application of spatial information systems, which has led to and increased the use of spatial data within multiple disciplines, which are not related most of the time to the purpose intended by a producer of the spatial dataset,
- the structures for accessing and interchanging data have to be fully documented in order to ensure that users (and computers) will understand the data,
- the interface between systems needs to be defined with respect to the data and operations using standardized methods – if the

structure of a dataset is standardized, the quality of the data can be easily defined and followed,
- the decreasing level of the quality of the data coming from its misuse is avoided.

The definition of the quality of spatial data helps users with their decision-making as to whether the dataset meets their criteria. From a producers' point of view, if the quality of their product is defined, evaluated and documented, this enables them to be successful in a market. A quality evaluation can be an expensive procedure. Although it is valuable, the costs must be weighed against the benefits gained from the information on the quality – the amount expended on the quality evaluation must be in reasonable proportion to the consequences of the errors discovered. Also, the willingness of the users to pay for a spatial quality evaluation (i.e., for the quality data) should be taken into consideration when justifying the level of the quality evaluation. Larry English, an authority on data quality issues, says that the business of non-quality data, including non-recoverable costs, the reworking of products and services, workarounds, and lost and missing revenue may be as high as 10-25% of the revenue or total budget of an organization (English, 1998).

UNIVERSE OF DISCOURSE

When the knowledge about a certain reality is represented in declarative language, the set of objects that can be represented is called a Universe of Discourse (UoD). The international standard *ISO 19101:2002 Geographic information – Reference model* defines a UoD as "the view on a real or hypothetical world that includes everything of the interest". UoD is an abstract model of a dataset, including only objects of interest, which can be viewed as similar to a "true value" in observation theory. The creation of UoD is sketched on a Fig.1

Any description of reality is always partial and just one of the possible interpretations of the real world. The creation of a UoD consists of following steps:

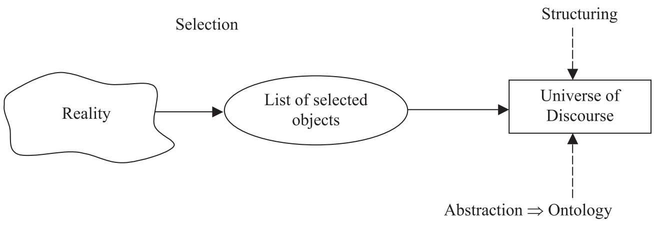


Fig. 1 *Creation of a Universe of Discourse (following Aalders, 2000)*

- the selection of objects of interest from the real world into the UoD,
- the abstraction of selected objects describing entities, the entity attributes and the relationships among the entities,
- the structuring of the entities and relationships.

The abstraction and structuring of selected objects results in an ontology used by dataset. "Ontology" is a term borrowed from philosophy, which defines it as a systematic account of existence. In geo-information and communication technology (geo-ICT) science the term "ontology" describes a structured, limitative collection of unambiguously defined concepts (Uitermark, 2001). The geo-ICT definition of ontology contains four items:

- an ontology is a collection of concepts instead of terms,
- the concepts are to be unambiguously defined,
- the collection is limitative and
- the collection has a structure, which means that the ontology contains the relationships between the concepts.

In order to ensure the stated level of the data quality, any concepts not defined in the ontology cannot be used.

Most UoDs are defined by glossaries that explain the meaning of each term used in the data definitions. This kind of definition can sometimes run into difficulties in its realization. If the UoD is defined as a set of objects supplemented by formal rules (which limit its possible interpretation), then the definition of the concepts is clear, and the realization of a UoD (e.g., by the creation of a dataset) appears to be easier. We can understand ontology as a synonym for the data definition. The concept of data definition is more familiar to the community of geographic information users. The ontology should follow a standardized method and can occur at different levels, i.e., international, national, domain and application levels. The ontology at a lower level should follow the definitions at a higher level, and the transfer of the meaning of the data definitions should be applied in different spatial datasets (Moellering et al., 2005). The ontology can be represented by taxonomies, node-trees, catalogues, dictionaries (formal definitions), thesauri, axioms, theorems or glossaries. When constructing spatial information systems, there are three main areas where ontologies are important (Frank, 2005):

- integration of data and interoperability,
- user interface design,
- setting the price for the information.

The specification of the UoD is performed after its definition, which means that objects from reality are selected and abstracted into the UoD. The specification of the UoD is performed for the application of the dependent selection of the objects as well as for data acquisition. The specification of the UoD (= ontology of the dataset) can be understood as a *conceptual schema of the data* (similar to a language's grammar) and *content description* (comparable to a language's vocabulary). The specification process is divided into two steps:

- *semantic specification* – specifies the description of the objects. It defines the objects and their attributes, attribute values, the relationships between the objects, the methods and inheritances.
- *geometric specification* – specifies the shape and absolute or relative position of the objects.

Other types of specification, e.g., the *input specification* (specifies the input method, idealization, generalization, etc.) can be included.

STANDARDIZATION OF THE ONTOLOGY

Feature catalogues (one of the possible representations of an ontology) form a repository for a set of definitions to classify real world phenomena of significance of a particular application field and its UoD. The catalogue provides a means for organizing the data, which represents these phenomena into categories so that the resulting information is as unambiguous, comprehensible and useful as possible. Fig. 2 describes the process from the UoD into a dataset as it is defined in the ISO 19100 series of standards.

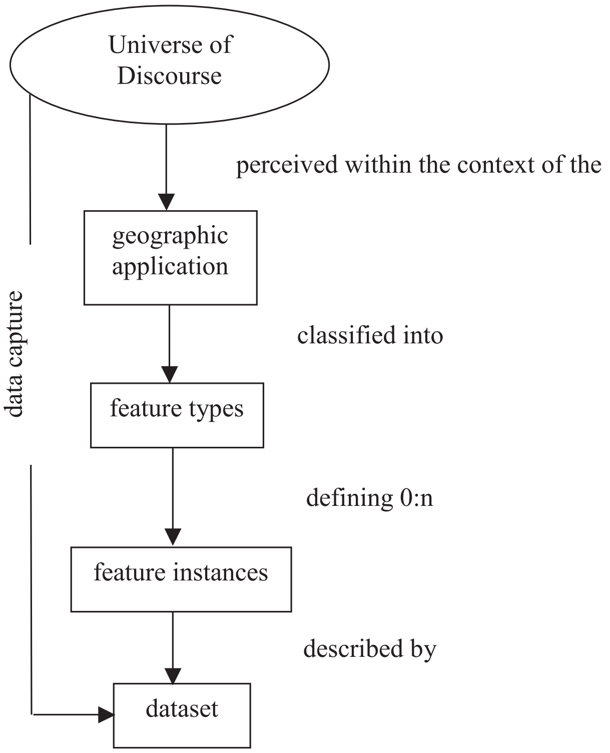


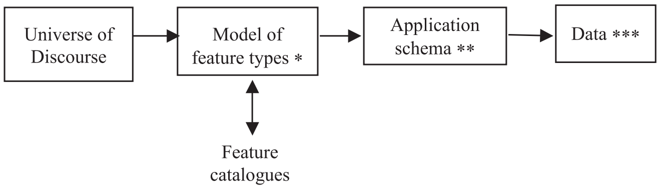
Fig. 2 *Process from the Universe of Discourse into a dataset (after ISO 19109: 2005 Geographic information – Rules for application schema)*

The availability of standardized feature catalogues that can be used many times will reduce the costs of data acquisition and simplify the process of product specification for geographic datasets. Geographical features occur at three levels: instances, types and behaviors. At the instance level, a single or group of features is represented as a discrete phenomenon, which is associated with its geographical and temporal coordinates and may be portrayed by a particular graphic symbol. These individual feature instances are grouped into classes with common characteristics – feature types. It is recognized that geographic information is subjectively perceived and that its content depends upon the needs of the particular application. The needs of a particular application determine the way instances are grouped into types within a particular classification scheme – the so-called "application schema". As the ISO/TC 211 (the creator of the ISO 19100 series of standards for geographic information) defines it, the purpose of the application schema is:

- to provide a computer-readable data description defining the data structure, which makes it possible to apply automated mechanisms for data management,
- to achieve a common and correct understanding of the data by documenting the data content of a particular application field, thereby making it possible to unambiguously retrieve information from the data.

The application schema defines the logical structure of the data and may define operations that can be performed on or with the data – it addresses the logical organization of a dataset instead of the physical.

The international standard *ISO 19109:2005 Geographic information – Rules for application schema* (ISO 19109) defines the rules for creating application schemas in a consistent manner to facilitate the acquisition, analysis, access, presentation and transfer of



** - model of the Universe of Discourse in terms of the concepts of the General Feature Model,*

*** - model of the structure and content of data in terms of a conceptual schema language,*

**** - data with a logical structure according to the application schema.*

Fig. 3 *Definition of feature catalogues (after ISO 19109: 2005 Geographic information – Rules for application schema)*

geographical data between different users, systems and locations. The creation of the application schema is a process, and its content according to the selected UoD has to be set. The content is modelled in terms of the feature types and their properties. When defining features, the following items should be described:

- definitions or descriptions used to group features into types,
- attributes associated with each type,
- relationships among the types,
- behaviour of the features.

To build an application schema, the ISO 19109 uses the General Feature Model (GFM) concept, which is a model of the concept required to classify an interpretation of the real world (i.e., the GFM is a metamodel of the feature types).

When describing the UoD as a spatial dataset using an application schema, it is necessary to respect the rules for reporting the data quality information. These are as follows (ISO 19109):

- the quality attribute shall be defined in an application schema and shall be used to carry quality information,
- the quality information of individual feature instances, their attributes, relationships and behaviour shall be reported by attribution whenever the quality of an instance is known to differ from its implied quality for the dataset or its part,
- the quality attribute shall be represented in the application schema as an attribute or a class that represents the data instance it reports,
- when complying with the ISO 19100 series, the quality attribute types shall be in accordance with the defined data types or sub-types in related standards, such as *ISO 19115:2003 Geographic information – Metadata* (ISO 19115) or *ISO 19113:2002 Geographic information – Quality principles* (ISO 19113),
- the quality attribute shall be documented in a metadata report (e.g., following ISO 19115).

STRUCTURE OF A SPATIAL DATASET AND QUALITY DESCRIPTION

A spatial dataset can be viewed as containing smaller groups of data, which have an identical feature type, feature attribute, feature relationship and behavior or sharing the same collection criteria or geographic extent. The reporting group can be as small as a feature instance, attribute value or occurrence of a feature relationship. The quality is defined for all levels of a dataset. Smaller groupings of data, which share expected commonality, have a similar quality. The data quality concepts shall allow for the additional reporting of the differing quality of the reporting groups together, and this approach provide a more complete picture of the dataset's quality. The quality information has its own quality (Aalders, 2000):

- confidence of the quality information,
- reliability of the quality information,
- description of the methodology used to derive the quality information,
- abstraction effect to account for the differences between the UoD and reality.

The ISO 19100 series of standards for geographic information recommends that the data quality requirements should be based on the General Data Quality Model, which shall be a part of the Data Product Specification (DPS) (the product in our terms is represented by a spatial dataset). The quality model is a set of quality parameters, and their measures are to be used to measure the quality of a spatial dataset and compare it to the reference dataset. The quality model must (Aalders, 2002):

- promote understanding both for the producers and users,
- be flexible and allow for all sorts of spatial data,
- be extensible and allow for new types of data,
- be practical,
- have a theoretical base (for future requirements).

QUALITY COMPONENTS OF SPATIAL DATA

Most of the experts on data quality (not only on spatial data quality) cite several attributes that collectively characterize the quality of data (Mayberry, 2002):

- **Accuracy** – does the data accurately represent reality or a verifiable source?
- **Integrity** – is the structure of the data and the relationships among the entities and attributes maintained consistently?
- **Consistency** – are the data elements consistently defined and understood?
- **Completeness** – is all the necessary data present?
- **Validity** – do the data values fall within the acceptable ranges defined by the business?
- **Timeliness** – is the data available when needed?
- **Accessibility** – is the data easily accessible, understandable and usable?

Spatial data are data about real objects and phenomena which may have spatial temporal and thematic components. From a geodetic and cartographic point of view a spatial component seems to be the most important characteristic of the spatial data. But spatial data in their role in the process of spatial modeling are not only about space, but also about the theme and time. It is true that without space, there is nothing spatial about the spatial data, but on the other hand, without a theme, there is only geometry. We can view space as a framework in which the theme and time is measured. Each of the dimensions of spatial data (space, theme and time) has some of the same components, such as accuracy, consistency and completeness.

To help users resolve the differences in the occurrence of the same feature from diverse sources, it is necessary to define the quality information of the spatial data according to the following steps:

- definition of the spatial data quality elements,
- derivation of easily understandable indices of the spatial data quality, which may accompany a dataset,
- setting up the methods for the representation or rendering of the specified data quality in its visualization.

TECHNICAL QUALITY PRINCIPLES

Various sources define different spatial data quality parameters. The Spatial Data Transfer Standard (SDTS) produced by the Federal Geographic Data Committee (FGDC) in 1992 defines the source, resolution, metric accuracy, thematic accuracy, completeness and logical consistency. The International Cartographic Association's (ICA) definition of data quality parameters adds semantic and temporal accuracy. In the European norms the homogeneity, usage and purpose occurs. After an evaluation of all the existing standards for data quality, the international standard ISO 19113 was (and up to now still is) fundamental for the categorization of data quality parameters bellow, but not only parameters highlighted in ISO 19113 are mentioned.

As Fig. 4 shows, two main types of data quality parameters are recognized – qualitative and non-qualitative. The important point is that ISO 19113 allows the producer of a spatial dataset to define the additional data quality parameters, so that after the evaluation of the user's requirements, sufficient data quality information can be produced and published.

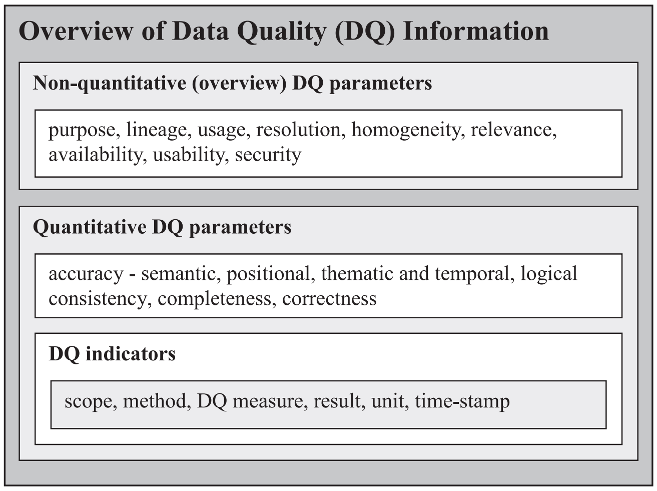

Overview of Data Quality (DQ) Information
Non-quantitative (overview) DQ parameters
purpose, lineage, usage, resolution, homogeneity, relevance, availability, usability, security
Quantitative DQ parameters
accuracy - semantic, positional, thematic and temporal, logical consistency, completeness, correctness
DQ indicators
scope, method, DQ measure, result, unit, time-stamp

Fig. 4 *Overview of data quality information (following the ISO 19100 series)*

Quantitative data quality parameters such are:

- **positional accuracy** – this is the expected difference between the position of the object from the dataset and its 'real' position and dimension in the real world (Veregin, 1999). The 'real' position of the object could be the more precise measurement – e.g., the position of the object obtained by geodetic measurement.
- **thematic accuracy** – this describes how well the thematic attributes are defined.
- **temporal accuracy** – this refers to a coincidence between the temporal coordinates of the object in the dataset and in reality.
- **semantic accuracy** – this is defined as the quality with which geographical objects are described in accordance with the selected model. Related to the meanings of the 'things‘ of the UoD, semantic accuracy refers to the relevance of the meaning of the geographical objects rather than the geometrical representation (Salgé, 1995).
- **completeness** – a comparison of the dataset with its product specification. The selection criteria are an essential determination of the completeness in geo-modelling of which the basis is the abstraction and generalization of the real world. There are two types of completeness there (Veregin, 1999):
    - ***data completeness*** – this describes the measurable balance of an omission detected in the dataset according to its product specification. Data completeness expresses a quality of the dataset, which depends on a particular application.
    - ***model completeness*** – this describes the level of coincidence between a dataset specification and the real world. Model completeness is dependent on the purpose and application for which a dataset is designed, so s review of all the particular use of the dataset is needed.
- **logical consistency** – this defines the degree of compliance of the logical relationship of a dataset's features. There must also be an evaluation of the logical consistency of geometrical and non-geometrical elements.
- **correctness** – the representation of a reality is correct, when the operations in reality have results which correspond to the results of the corresponding operations in the representation.

Non-quantitative (overview) data quality parameters are:

- **purpose** – this describes the rationale for creating a dataset and contains information about its intended use.
- **usage** – this describes the application(s) for which a dataset has been used.

- **lineage** – this describes the history of a dataset and recounts the life cycle of a dataset from the collection and acquisition throughout the compilation and derivation to its current form (ISO 19113).
- **resolution** – defines the smallest object in terms of the space, time or theme, which can be recognized in the dataset.
- **homogeneity** – is the textual and qualitative description of the expected or verified unity of the qualitative parameters in a dataset (STN P ENV 12 656, 2001). It refers to how valid the statements about the quality are.
- **relevance** – this could inform a user when the dataset was created. This information could be very important for the datasets containing data from the dynamic extent (extent with frequent change).
- **availability** – this could inform the user of the author's right.
- **usability** – links the information back to the real world.
- **security** – protects the data from any 'unsecured contact' in the sense of being reliable for the users (protected data are much more reliable than data which anyone can 'touch').

The data quality elements have its data quality sub-elements, of which descriptors are:

- the data quality scope,
- the data quality measure,
- the data quality evaluation procedure,
- the data quality result,
- the data quality value type,
- the data quality time-stamp.

MODELING QUALITY IN THE SELECTED SPATIAL DOMAIN – THE SLOVAK CADASTRAL DOMAIN

In order to explain the theoretical data quality principles in a spatial domain, the Slovak Cadastral Domain (SCD) has been selected by the author. Data quality modeling in the SCD respects its constraints defined by related legal framework. This legal framework can be in seen in terms of the ISO 19100 series of standards as the only data product specification (which should determine the producer's expectation of the data quality parameters). The parameters of the spatial data quality as defined above are defined, assessed, evaluated or reported in this domain as follows:

- **Lineage** – the lineage of the cadastral domain is not reported. Users, who are educated or aware of the history of building the cadastre, can assess the lineage of this domain.
- **Usage** – the cadastral law (Act. No.162/1995 – §1 and §2) determines the usage of the cadastral data.
- **Purpose** – the cadastral law (Act. No.162/1995 – §1 and §2) determines the purpose of the cadastral data.
- **Resolution** – the cadastral map is the geometric representation of the cadastral data. The scale of the cadastral map (1:1000, 1:2000 or 1:5000) and the regulations related to the production of the cadastral map determine the geometric resolution of the cadastral data. The thematic resolution is determined by Act. No.162/1995 – §6, which defines the subject of the cadastre – the objects concerned. The temporal resolution is not specifically determined, but can be derived from the Act. No.162/1995 as follows: the object in the cadastral domain exists, until the opposite is not proved.
- **Positional accuracy** – this data quality parameter is the only one, which is clearly defied and stated in the data product specification by the list of data quality codes giving a quality label to each spatial object. The quality label from the list of quality codes represents the definition of the positional accuracy of every object in the cadastral domain model. This part of the quality analysis of the cadastral domain of Slovak Republic is documented in the quality model of the SCD – for details, see (Ivánová, 2006).
- **Attribute accuracy** – the definition of attribute accuracy is absent in the specification of the Slovak cadastral domain. This is determined only by the definition of the attribute features (present in the Set of Descriptive Information about the estates) as a definition of the types of information to capture.
- **Temporal accuracy** – is determined by the updating policy of the cadastre.
- **Logical consistency** – the definition of the logical consistency is lacking. Only the consistency between the types of representation of a certain object (geometric and 'descriptive') is maintained – through the index (= identification number) of the object, which is represented by the number of the parcel.
- **Completeness** – the definition of the completeness is absent in the specification of the cadastral domain of the Slovak Republic.
- **Semantic accuracy** – the detailed specification of the semantic accuracy is lacking in the specification of the Slovak cadastral domain. But if we concern the producer's specification – determined by the legal constraints, it is obvious that the semantic accuracy is well preserved by fulfilling the prescriptions for the inputs of the objects to the cadastral dataset.
- **Correctness** – this parameter is determined by the general rule for the cadastre, which is that the information in the cadastre is correct (relevant, accurate, valid, etc.) if the opposite is not proved.
- **Usability, homogeneity, relevance and availability** – the definition in the specification is lacking and the assessment of these parameters is done only by the 'inherited experience' (to ask other users about their experience).

The evaluation of the quality of the Slovak cadastral domain is performed by accident – if anyone finds a mistake (in whichever sense) amongst the cadastral data the responsible person is obliged to correct the mistake. The correction is coordinated by the technical regulations for the correction of mistakes in the cadastral dataset. The documentation of the quality (but also the cadastral data itself) is not maintained by any regulation.

If the user of the cadastral data wishes to know about the parameters of the data quality, he is obliged to study the set of regulations (representing the legal constraints of the cadastral domain) or simply trust the reliability of the purchased data.

CONCLUSION

Expert on quality (who is not else as manager) keeps on explaining what quality means in a certain organization instead of making formulas and producing documents – it is very difficult to talk about spatial data quality in general terms.

Within the context of a theory of spatial data quality, the following general requirements for the application of the principles of data quality concepts in practice arose from the research behind the article:

- It is necessary to change the complex view on a spatial dataset through the entire process of its creation – to be able to set a proper data quality model;
- It is necessary to strengthen the data product specification – in the event the data product specification does not exist, this could mean that the idea of the dataset's purpose is not very clear; therefore, it is very difficult to talk about quality and to answer the basic questions:
  - Who are the users?
  - Which data quality parameters are crucial for the user?
  - Which data quality parameters are crucial for producer?
  - What data quality parameters should be documented?

The use of standards, especially those related to quality of spatial information together with the set of standards for quality management system would increase the ability of the organization handle spatial data quality. The presented article gave a brief introduction to understanding of the quality in the spatial dataset.

aknowledgment

The research presented in this article is a part of the supported by the VEGA under the contract No. 1/4025/07

REFERENCES

- **Aalders, H.J.G.L.** (2000): *The universe of discourse in GIS*. In: Quality of Geo-Information Course Material, GIS Section, Faculty of Geodesy, 2002/2003, TU Delft, The Netherlands
- **Aalders, H.J.G.L.** (2002): *The Registration of Quality in a GIS*, In: **Shi, W., et al.** (2002): *Spatial Data Quality*, Taylor & Francis, London, UK, ISBN 0-415-25835-9, pp. 186-200
- **English, L.** (1998): *The High Cost of Low-Quality Data*, DM Review Magazine, January/1998, (http://www.dmreview.com/article_sub.cfm?articleId=771) (8.6.2006)
- **Frank A.** (2005): Ontology for GIS (Draft – v4), http://www.geoinfo.tuwien.ac.at/index.php - (downloaded 1.6.2005)
- **Ivánová, I.** (2006): *Data quality in spatial datasets*, PhD. thesis, Department of Theoretical Geodesy, Faculty of Civil Engineering, Slovak University of Technology, Bratislava
- **Mayberry, M.** (2002): *Data Quality: Before The Map is Produced*, http://www.directionsmag.com/article.php?article_id=250 (8.6.2006)
- **Moellering, H., Aalders, H.J.G.L, Crane, A.** (2005) – World Spatial Metadata Standards, © ICA, Elsevier Science/Pergamon USA, ISBN – 0080439497
- **Salgč, F.** (1995): *Semantic accuracy*, In: **Guptill, S., Morrison, J.**, (1995): *Elements of Spatial Data Quality*, BPC Wheatons Ltd., Exeter, ISBN 008042432 pp.139-151
- **Uitermark, H. T.**, (2001): *Ontology-Based Geographic Data Set Integration*, Deventer, NL, ISBN 90-365-1617-X
- **Veregin, H,** (1999): *Data quality parameters*. In: **Longley, P.A., et al**: *Geographical Information Systems*, Second Edition, John Wiley and Sons, New York, pp. 177–189.
- **Act No. 162/1995** *Cadastral law* www.geodesy.gov.sk (8.6.2006)
- *ISO 19101: 2002 Geographic information – Reference model*, ISO, Switzerland, 2002
- *ISO 19109:2005 Geographic information – Rules for application schema*, ISO, Switzerland, 2005
- *ISO 19113:2002 Geographic information – Data quality principles*, ISO, Switzerland, 2002
- *ISO 19115:2003 Geographic information – Metadata*, ISO, Switzerland, 2003
- *STN P ENV 12 656:2001, Geografická informácia. Opis dát. Kvalita* (translated as: *Geoigraphic information. Data Description. Quality*), SÚTN, Slovakia, 2001