Data Quality Standards

Gerhard JOOS, Germany

Key words: quality of geospatial data, international standardization, quality related metadata

SUMMARY

The quality of the geodata is important information in order to judge the reliability of analysis results obtained in a GIS or via a web service. To be ale to compare quality information the results have to be obtained based on the same concept and principles using the same data quality measures. This is the motivation for standardizing the concepts of data quality in the geospatial community. Since geospatial information is multidisciplinary and not restricted to political borders, the standardization effort has to be based on international consensus. ISO/TC 211 is the international standardization body responsible for GIS standards. This committee has produced a series of international standards including standards dealing with quality. The author who was actively involved in the standardization process gives an overview of existing quality-related standards and how these standards interact.

ZUSAMMENFASSUNG

Die Qualität von Geodaten ist eine wichtige Information, um die Zuverlässigkeit von Analyseergebnissen abzuschätzen, die in einem GIS oder einem Webdienst auf der Basis von Geodaten ermittelt wurden. Um Informationen über die Datenqualität vergleichbar zu machen, ist es erforderlich, dass diese Angaben nach den gleichen Prinzipien und einem einheitlichen Konzept und mit identischen Qualitätsmaßen gemacht werden. Das ist der Grund warum der Bereich Datenqualität als Norm festgeschrieben werden muss. Da Geoinformatik eine multi- und interdisziplinäre Wissenschaft ist, die nicht an politischen Grenzen halt macht, muss die Entwicklung der Normen auf internationaler Ebene durchgeführt werden. Das ISO/TC 211 ist das internationale Normungsgremium, das für Geoinformation zuständig ist. Zu den Normen des ISO/TC 211 zählen auch eine Reihe von Normen, die Datenqualität zum Normungsgegenstand haben. Der Autor, der selbst bei der Entwicklung dieser Normen aktiv mitarbeitet, gibt einen Überblick über Normen, die sich mit Datenqualität beschäftigen und zeigt, wie diese einzelnen Normen zueinander passen.

Data Quality Standards

Gerhard JOOS, Germany

1. INTRODUCTION

Quality is an intrinsic property of any product. If a product fits to the specific requirements of a potential user depends on the intended use. Nonetheless there are products that even when their product specification fit the users needs are still not completely satisfactory or just useless to the user, because their actual constitution derivates from the product specification. If this derivation is unintended or random it affects the quality of the product.
In order to keep the constitution of products in a production process for series of product instances constant and predictable over time rigorous procedures to maintain or even improve production processes have been introduced and standardized. The international standards ISO 9000 ff are concerned with quality management and describe guidelines, terms and quality models (ISO 9000ff, 2000). The ISO 9000 family of international quality management standards and guidelines has earned a global reputation as the basis for establishing quality management systems. Particularly ISO 9001:2000 specifies requirements for a quality management system for any organization that needs to demonstrate its ability to consistently provide a product that meets customer and applicable regulatory requirements and aims to enhance customer satisfaction. The standard is used for certification and contractual purposes by organizations seeking recognition of their quality management system.

Geospatial data can be considered as products or as a series of products that are produced with respect to a product specification in order to satisfy the needs for a particular user or for a variety of users that intend to use the geospatial data for different applications. Since different applications require different constitutions of the data, it is not possible to state the fitness for use in a common way. Therefore it is required to state the quality of geospatial data in an objective way and since the quality of different datasets must be comparable the quality description has to be in a standardized form. ISO/TC 211 provides a series of standards that deal with various aspects of geographical information / geomatics which include in particular ISO 19113:2002 Quality principles, ISO 19114:2003 Quality evaluation procedures, ISO 19115:2003 Metadata and the technical specification ISO/TS 19138 Data quality measures that is about to be released. This paper will introduce the various standards and show their interdependencies and their application for data producers and data users.

2. QUALITY MANAGEMENT SYSTEMS

The term quality management is tightly coupled with the ISO 9000 series of standards (ISO 9000:2005, ISO 9001:2000, ISO 9004:2000). Just by looking at the principles behind ISO 9000 that are listed below, it can be deduced that ISO 9000ff are mainly focusing on the management aspect to achieve costumer's satisfaction.

- Principle 1 Customer focus
- Principle 2 Leadership
- Principle 3 Involvement of people
- Principle 4 Process approach
- Principle 5 System approach to management
- Principle 6 Continual improvement
- Principle 7 Factual approach to decision making
- Principle 8 Mutually beneficial supplier relationships

Also for the production of geodata – capturing and updating geospatial features – these principles have to be obeyed. When implementing a quality management system for geospatial information it gets obvious that important pieces are missing, e.g. that there have to be criteria to judge to quality of geospatial data (Jakobsson, 2002). These criteria have to be objective so that they can be applied independent of the intended application. They have to be standardized in order to make quality information comparable. This was envisioned when the international standardization committee for geographical information ISO/TC 211 was established. The program of work contains several work items that are related with quality aspects of geospatial information. The quality related standards that came out of the standardization work of ISO/TC 211 will be presented and discussed in the following sections.

3. QUALITY PRINCIPLES

3.1 Basic concept

The quality of a dataset shall be described in an objective and generic way. What should be expressed is, how close the information in an actual dataset represents the situation in the real world. Since the real world is too complex to comprehend as a whole let alone to capture in geospatial features, it is required to build a model. The model is driven by user requirements and formulated as a product specification. The product specification gives the reference of the content and the structure of the dataset which is called the universe of discourse. It is an abstract concept that represents the targeted dataset for capturing. It can be envisioned as the ideal though not existing dataset. The actual data has to get as close as possible to this universe of discourse. Any difference can be considered as error. The concept can be visualized as shown in Figure 1.

ISO 19113 provides a classification scheme for these errors. They are categorized into different elements and subelements depending on the nature of the error. This classification scheme is the most important content of this international standard.

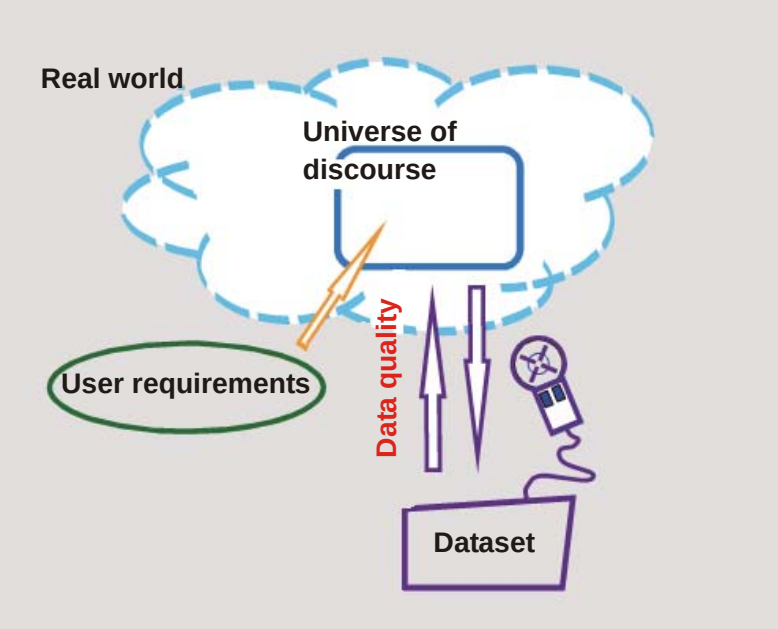

Figure 1: Concept behind data quality principles in ISO 19113

3.2 Classification Scheme

The standard on quality principles does not only provide the concept of data quality it also standardizes the names and a scheme under which all differences of a dataset and the corresponding universe of discourse can be categorized. The list of elements and subelements as described in ISO 191913 is shown in Table 1**.**

Table 1: Elements and subelements to categorize data quality aspects in ISO 19113

completeness – presence and absence of features, their attributes and relationships • commission – excess data present in a dataset • omission – data absent from a dataset
logical consistency – degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical) • conceptual consistency – adherence to rules of the conceptual schema • domain consistency – adherence of values to the value domains • format consistency – degree to which data is stored in accordance with the physical structure of the dataset • topological consistency – correctness of the explicitly encoded topological characteristics of a dataset
positional accuracy – accuracy of the position of features • absolute or external accuracy – closeness of reported coordinate values to values accepted as or being true • relative or internal accuracy – closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true • gridded data position accuracy – closeness of gridded data position values to values accepted as or being true

temporal accuracy – accuracy of the temporal attributes and temporal relationships of features • accuracy of a time measurement – correctness of the temporal references of an item (reporting of error in time measurement) • temporal consistency – correctness of ordered events or sequences, if reported • temporal validity – validity of data with respect to time
thematic accuracy – accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships • classification correctness – comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference dataset) • non-quantitative attribute correctness – correctness of non-quantitative attributes • quantitative attribute accuracy – accuracy of quantitative attributes

4. QUALITY EVALUATION PROCEDURES

Since ISO 19113 provides only the principles how geospatial data can be described it gives no guidance on assessing the quality of actual datasets. That's where ISO 19114 geographic information – quality evaluation procedures comes in. This standard provides a framework of procedures for determining and evaluating quality that is applicable to digital geographic datasets. This framework is consistent with the data quality principles defined in ISO 19113. Part of the scope of ISO 19114 is to establish a framework for evaluating and reporting data quality results either as part of data quality metadata only or also as a quality evaluation report.

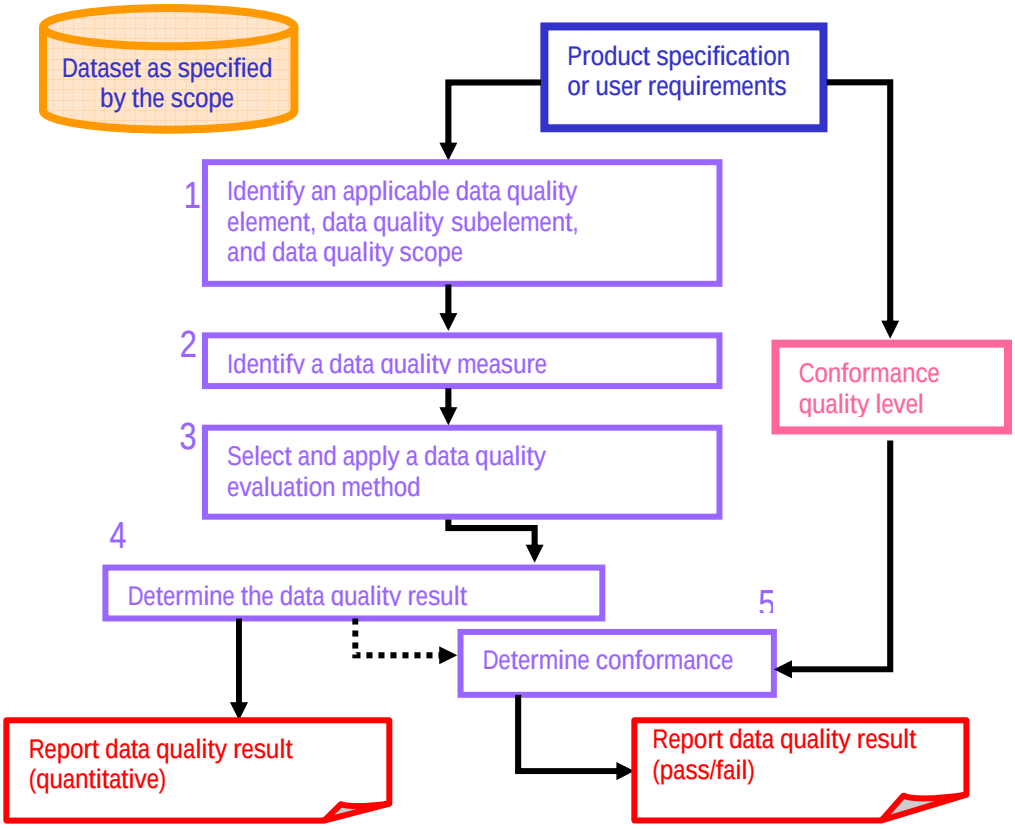

Figure 2: Five steps for quality evaluation according ISO 19114

The basis for all quality evaluations is the product specification since this specifies the universe of discourse which is the reference for the determination of data quality. How a product specification has to be written is determined by the ISO standard 19131.
A quality evaluator has to decide which parts of the dataset have to be evaluated (scope) and what aspects of the data quality (data quality element and subelement) should be evaluated. This part under evaluation is called data quality scope. Suitable evaluation methods that fit to the selected data quality measures have to be chosen. Step 2 shows the importance of data quality measures. The evaluation method gets applied in order to determine the data quality result, that can either be used to decide if the dataset fulfills certain quality requirements expressed as conformance quality level or to report the determined result as metadata or in a specific quality report. ISO 19114 gives annexes with examples on data quality evaluations.

5. DATA QUALITY MEASURES

As discussed in the previous clause quality evaluation results are expressed by data quality measures. Since 19113 just provides the principles and stops at the data quality subelement level the data quality measures have to be provided from different sources. In order to make data quality evaluation results comparable independent of their origin it is important to use standardized data quality measures. ISO 19138 gives a structure how data quality measures shall be built. It also provides data quality base measures that can be used in many cases to

construct data quality measures and as an normative annex it provides a list of data quality measures that is meant as an initial input for an online register that can be accessed to query data quality measures. The register is the preferred way of providing data quality measures since it can be used by a human user who is looking for a specific measure and by software alike. The software can reference to this register in order to avoid repetitive description of the same measure.

5.1 Data Quality Base Measures

There are principles that can be applied to most data quality measures. For example for an error classification of type Boolean it is possible to report whether something is either correct or wrong or it is possible to count all errors of the same category. The count result can then be used as measures for this certain aspect of data quality. ISO 19138 provides a set of counting related data quality base measures that can be used to build a data quality measure without unnecessary duplication. There are also data quality base measures for uncertainty related topics. They may be used to describe the uncertainty of arbitrary one or more-dimensional random variables. To construct actual data quality measures the random variables have to be specified for quantitative attribute value uncertainty or for positional accuracy. The base measures are used for the measures that are provided in the standard and they can be used to build new measures to fulfill the user's requirements. Anybody constructing new data quality measures for example in a maintenance process of a register has to check, if a data quality base measure exists, on which a new data quality measure can be based.

5.2 Examples of Data Quality Measures

Each data quality measure is described by 13 components. Some of them are optional. They are defined in conformance with the international standard for registration of geographic information items ISO 19135. Table 2 gives an example of a very simple data quality measure.

Table 2: Number of incorrectly classified features

Line	Component	Description
1	Name	number of incorrectly classified features
2	Alias	–
3	Data quality element	thematic accuracy
4	Data quality subelement	classification correctness
5	Data quality basic measure	Error count
6	Definition	number of incorrectly classified features
7	Description	–
8	Parameter	–
9	Data quality value type	Integer
10	Data quality value structure	–
11	Source reference	–
12	Example	–
13	Identifier	62

5.3 Register for Data Quality Measures

Data quality measures are by no means a closed set. As geoinformatics evolves new data quality measures will be required and designed by different user communities. To participate in the advantages of standardized data quality measures as discussed previously it is required that these new measures get included in the list of standardized measures. A printed standard gets only reviewed every 5 years. This period is too long for distributing new measures to the community that requires it. A web tool that serves as a repository can solve this problem. The ISO standard 19135 gives all the organizational requirements to set up and maintain such a register. A registration authority will be responsible for new data quality measured to be added to the register, or existing ones changed or deprecated.

6. METADATA

Metadata are important in many ways. They are for example required to document the origin of geodata, to give a potential user hints on the usage and limitations for using the data and for discovering data in a catalog service among other application fields. In that sense is the quality information an essential part of metadata. For that reason the international standard on metadata for geospatial data – ISO 19115 metadata – contains parts dealing with data quality. ISO 19115 contains the only UML diagram that models the concepts of data quality. Metadata elements are defined to report quality evaluation results.

7. CONCLUSIONS

Standardization is the way to achieve interoperability. This is not only true for GIS, for data and services access on the web, but also for metadata and in particular data quality. It has been shown that there exist a number of quality-related standards as part of the ISO 19100 series of standards developed by ISO/TC 211. The different functions of the standards and their differentiations were discussed also with taking the ISO 9000 quality management standards into account. The slight inconsistencies in minor technical aspects between the listed standards were not discussed although they exist. ISO/TC 211 is aware of this open task and has started efforts to harmonize the listed standards.
The advantages of a publicly available register for data quality measures have been discussed. The implementation and publication of such a register for use by the geo-community is also an open issue.

REFERENCES

Chrisman, N. R., 1991: The error component in spatial data. In: Maguire D J, Goodchild M F and Rhind D W (Eds) Geographical information systems, Wiley, New York, pp. 165-174

Goodchild, M.F., 1991: Issues of quality and uncertainty. In: Advances in Cartography, J. C. Muller (Ed), vol. 6, pp. 113-139

Jakobsson, A., 2002: Data quality and quality management – examples of quality evaluation procedures and quality management in European national mapping agencies. In: Spatial data quality, Wenzhong Shi, P. F. Fisher, and M. F. Goodchild (Eds.), Taylor & Francis, London, New York, pp. 216-229

Joos, G., 2003: Standardization of Data quality measures. In: Proceedings of the second International Symposium on Spatial Data Quality, Wenzhong Shi, M. F. Goodchild, and P. F. Fisher (Eds.), The Hong Kong Polytechnic University. pp. 205-209

International Standards

ISO 9000:2005, Quality management systems -- Fundamentals and vocabulary
ISO 9001:2000, Quality management systems -- Requirements
ISO 9004:2000, Quality management systems -- Guidelines for performance improvements
ISO 19113:2002, Geographic information -- Quality principles
ISO 19114:2003, Geographic information -- Quality evaluation procedures
ISO 19114:2003/Cor. 1:2005, Geographic information -- Quality evaluation procedures – Corrigendum 1
ISO 19115:2003 19115, Geographic information -- Metadata
ISO 19115:2003/Cor. 1:2005, Geographic information -- Metadata - Corrigendum 1
ISO 19131:2006, Geographic information – Product specification
ISO 19135:2005, Geographic information -- Procedures for item registration
ISO 19138:2006 (to be published), Geographic information – Data quality measures

BIOGRAPHICAL NOTES

Dr. Gerhard Joos studied surveying engineering at the University of Stuttgart in Germany and at the University of Calgary in Canada. After his graduation with an engineering diploma he started as research fellow for the subject GIS at the University of the Bundeswehr in Munich, Germany. In 1999 he finished his Ph.D. on the topic “quality of object-structured geospatial data”. Dr. Gerhard Joos participates in standardization of geoinformation since 1998. He is active member of the Open Geospatial Consortium, the ISO/TC 211 and the Digital Geographic Information Working Group (DGIWG). He leaded the project on developing an international standard of data quality measures that is just about to be released as ISO 19138.

Dr. Gerhard Joos is currently working as a freelance consultant for geospatial information. He recently received an offer from the Danish Technical University as associate professor for geoinformation.

CONTACT

Dr. Gerhard Joos
dotGIS – consultancy for geospatial intelligence
Riedhauser Str. 2b
85649 Brunnthal
GERMANY
Tel. + 49 8102 773719
Fax + 49 8102 773818
Email: Gerhard.Joos@dotGIS.de
Web site: www.dotGIS.de