UDC528.001 + 681.518

FUNCTIONAL MODEL OF AUTOMATED SYSTEM FOR GEOSPATIAL DATA QUALITY CONTROL AND EVALUATION

A. Lyashchenko, M. Gorkovchuk
Scientific Research Institute of Geodesy and Cartography,
Kyiv National University of Construction and Architecture

Keywords: geospatial data, geospatial databases, data quality, spatial data quality evaluation.

Problem definition

Objectively, that data quality is very important factor for the operation of any information system. Even the best mathematical decision-making models, realized by software information systems, practically lose their meaning at low or unknown data quality, since the reliability of simulation results will be low or unknown, which negates the very idea of computer information technology in any field of activities.

Quality evaluation of topographic and geodetic survey, in particular coordinate accuracy of objects, belongs to the classic problems of traditional topographic-geodesic production. But transition from cartographic to digital (geoinformation) models requires both clarification of «quality» term for new final topographic-geodesic product in the form of spatial datasets and databases as well as the development of principles, methods, procedures and technologies of product quality control and evaluation.

The structure and content of spatial datasets and databases usually differs from traditional digital topographic maps. They are characterized by specific organization scheme, object composition, set of attributes, relations and data integrity constraints, defined by technical requirements for the data model, depending on purpose, functions and tasks on applicable geoinformation system (GIS). Therefore, the quality of geospatial data defined as the set of characteristics that reflect the ability to meet established or anticipated needs of applications.

Considering that that the actual GIS integrates data from different origins, and their amounts are significant, since it deals with spatial characteristics and other information about the tens of thousands (or even millions) of real-world objects, efficient quality assurance of geospatial data is possible only through the use of specialized tools for automated support and compliance control of implementation data models with certain applied requirements.

Analysis of the researches and publications, related to problem solution

In recent years the publications devoted to the problem of geospatial data quality highlights the results of reseaches in a following areas:

1. Development of data quality concept, unification of principles and models of complex spatial data quality management system [6 – 9, 12].

2. Development and implementation of international data quality standards, which are a set of standards ISO 19100: Geographic information/Geomatics. This include international standard ISO 19115 Geographic Information - Metadata and the draft standard ISO/DIS 19157:2011 Geographic information - Data quality [4], which review and summarize the statements of standards ISO 19113 Quality principles, ISO 19114 Quality evaluation methods and ISO/CD TS 19138 Data quality measures, and which is going to replace these standards.

3. Justification and unification of measures and procedures of quality evaluation. Reliability of quality assessment largely depends on adequacy of choice of these measures and procedures.

4. Creation of tools for automated quality control and evaluation. Developers of instrumental GIS and spatial database management systems (SDBMS) try to integrate instruments for spatial data quality control and assurance in their products. For example, subsystem Data Reviewer is realized as a part of ArcGIS for automated spatial data quality control [13]; starting from version 10 Oracle Spatial is supplied with Topology and Network Data Model applications, which allow to control topological consistency according to the predefined rules [10]; the library of functions for control of geometrical models and topology of spatial data in DBMS PostgreSQL+PostGIS is constantly expanding [11].

Generalization of the main research results in these areas and justification of conceptual principles of organization of spatial data quality assurance and evaluation system is reflected in a previous publication [1].

Prepared by the EuroGeographics's experts Guidance on implementation of basic quality standards ISO 19100 in national cartographic and cadastral agencies (NCCA) [7] and other works pay attention to the growing possibilities of the Internet to obtain geospatial data that are created for different purposes and different manufacturers. In these conditions the information about different objects in the databases can have different accuracy, measure periods and more. These features in modern NCCA activities experts named "new paradigm of production", which involves multiple sources of spatial data in partnership with the survey departments of various agencies ("production of many types"). This new

paradigm enhance the role and objective needs in methods and means of achieving interoperability of geographic information application models, as well as uniform procedures and services of provision and assessment of data quality.

Task definition

Proposed work justify the basic principles of construction, structure and functions of automated system for complex challenges of assurance and evaluation of spatial data quality for all stages of their lifecycle based on general concept of geographic information quality, as defined in ISO 19100 set of standards [4].

Account of the main material

It is known, that quality evaluation on a final stage of production gives inadequate result. It cannot efficiently influence the improvement of product quality, so continuous quality control system on all production stages is necessary. Continuous control covers the whole state of production and is defined by ISO 8402 as “Full quality”. Implementation of this approach requires the use of a Quality Management System (QMS), particularly definition of permissions, responsibilities and interactions between all participants and all processes and procedures, that are important for achieving quality objectives of the organization and conducting long-term sustainable quality policy.

Geospatial data quality assurance system can be defined as a set of organizational and technological measures, methods and specific software for quality control at all stages of collection, input, processing and use of geospatial data, which are recorded, collected and updated in geospatial database and exported to external systems as spatial datasets and/or digital maps. Quality control is integral part of business process for any method of collecting, input, recording and exporting spatial data.

ISO/DIS 19157:2011 define quality of spatial data as a level of correspondence of a data to formulated requirements.

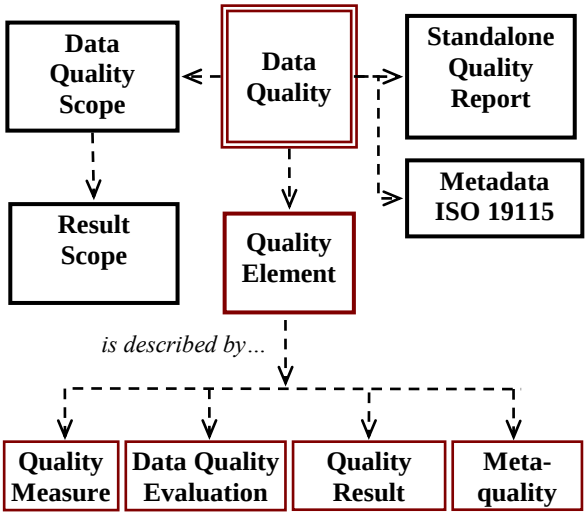

Fig. 1. Concept of spatial data quality model [4]

According to ISO 19157 conceptual model (fig. 1) quality describes spatial data characteristics, which make up quality scope – complex of datasets, standalone dataset or a part of dataset, which is characterized by uniform quality requirements.

Spatial data quality consist of descriptive and quantitative quality elements. Each quality element is characterized by quality measure, method and result of evaluation, and a set of descriptive quality elements in metadata. Results of evaluation form corresponding report and chapter of quality metadata, which structure and content is described in ISO 19115. Usually quality report is using for correcting the detected errors, and the metadata are attaching to spatial dataset as a final product and can be used by consumers for evaluation of external or applied quality.

According to ISO 19157 spatial dataset quality consist of descriptive and quantitative elements.

Descriptive quality elements provide common nonquantitative quality information. They allow further evaluation of the usefulness of the data set for a specific application, contain information about its purpose, use and origin. Elements of "Purpose" and "Use" describes the scope of the dataset. Use of the dataset is determined by the developer or data user. Element "Origin" describes the history of the formation of the dataset and to some extent its life cycle, beginning with the collection process, the following coding and conversion to the current data format.

Quantitative quality elements of spatial data make it possible to assess whether a particular dataset meets the criteria stated in the specifications of the product. Aspects of data quality elements are called subelements. Subelements evaluated or tested in various ways. According to the general concept of data quality assessment, not all elements and sub-elements, and not all means of their assessment and verification can fit a particular dataset. In addition, some sub-elements can be applied to the entire dataset, can be evaluated or tested for it, while others apply to smaller groups of data belonging to the dataset of higher level and can be evaluated or tested for them.

ISO 19113 determines such basic quantitative quality elements and subelements:

completeness (with commission and omission subelements);

logical consistency (conceptual, domain, format and topology consistency);

positional accuracy (absolute or external accuracy, relative or internal accuracy and positional accuracy of raster data);

thematic accuracy (correctness of classification, the correctness of nonnumeric attributes, the accuracy of numeric attributes);

temporal accuracy (precision of time measurement, time consistency, time conformity).

Elements of data quality are determined as means of direct detection and the development of different kinds of information about quality. Typically, data quality sub-elements are interconnected. For example, the error in the coordinates may cause at least two types of errors - positional and topological.

Quality measure is quantitative characteristic of elements (subelements) of spatial data quality. Unification quality measures is carried out to achieve compatibility and comparability of quantitative information about the quality of different datasets. One of the main requirements for quality measures is unambiguous of their determination and validity of calculation methods.

The draft ISO/DIS 19157:2011 proposes a set of standardized measures of quality that allow to evaluate almost all quantitative elements and sub-elements of spatial data quality. Basic quality measures are grouped into two classes:

1) measures based on a counting of erroneous or correct objects;

2) measures based on modeling uncertainty by statistical methods.

First class has six basic quality measures, which are based on different methods of evaluation of number of correct or incorrect objects, namely: error indicator (yes or no), correctness indicator, error number, number of correct objects, error percentage and percentage of correct objects.

Numerical data obtained from measurements with a certain precision and therefore it is recommended to use statistical methods for assessment the uncertainty of some of the measured values.

The international standard ISO 19138 and the draft ISO/DIS 19157 use basic quality measures specified for all quantitative elements and sub-elements of spatial data quality, which ultimately allowed the identification of more than 80 individual quality measures. However, due to the specificity of data quality, this list may not be comprehensive. With the development of GIS spatial data quality requirements are increasing, so it is a natural to develop additional quality measures.

Conceptually, most researchers distinguish the internal and external quality of spatial data [6,9].

The internal quality is understanding as a level of correspondence between the created and the "ideal" spatial dataset, which would have been made (ie, the data produced without errors). This set is a reflection of the real world on a certain date in accordance with the specifications of products, establishing a set of rules and requirements of the transition from the real world into the data model. Specifications include, for example, the list of objects to be displayed, the type of geometry for each class of objects, attributes that describe the objects and allowable values (domains) for attributes. The internal quality can be described in various ways, but mainly criteria defined by the International Standard ISO 19113 are used, namely completeness, logical consistency, positional, thematic and temporal accuracy. A detailed description of the data model of objects, attributes and relations between objects is presented as a catalog of classes of geospatial objects according to the rules defined in ISO 19110 Geographic information - Methodology for cataloging objects [2, 3].

External quality refers to level of compliance of the finished product to needs or expectations of the user.

This quality is not absolute, and therefore the same set of spatial data may have different external quality for different users. External quality is often defined as the suitability of the product for use.

Figure 2 represent the functional scheme of proposed complex system for automated quality control and evaluation of spatial data (the system) during their production at all stages of the lifecycle of geospatial data: preproduction, production, use, maintenance, renewal. The system has the base-oriented architecture, which provides that all information system resources are stored and maintained in an environment of universal DBMS.

Modern object-relational databases are extended with software tools for processing and analyzing of geospatial data in the environment of DBMS at the level of access according to the Structured Query Language standard SQL 3 [5], for instance, Spatial Option in Oracle [10] or PostGIS in PostqreSQL [11]. It allows to store and maintain created and/or estimated spatial datasets in a universal object-relational database. Spatial database also record metadata (common metadata about spatial datasets, metadata about spatial properties and attributes of dataset objects) at all stages of data production. External datasets stored in instrumental GIS format can be imported in the system and loaded into DBMS for control and evaluation of their quality using data quality knowledge base. The system consist of several functional subsystem, which provide:

– import/export with format conversion;

– support the processes of creation data specifications with a feature catalog, logical diagram of the data model, the digital description rules and topology requirements for the elements/sub-elements, measures and methods for assessing the quality of data that are stored in a formalized form in feature catalog databases, internal and external (applied) quality models;

– maintaining a knowledge base for the control and evaluation of data quality;

– monitoring and evaluation of data quality by quality elements that characterize logical consistency, positional, thematic and temporal accuracy;

– formation and documenting reports on the results of control and evaluating the quality of spatial data, including metadata sections with information on the quality of spatial datasets.

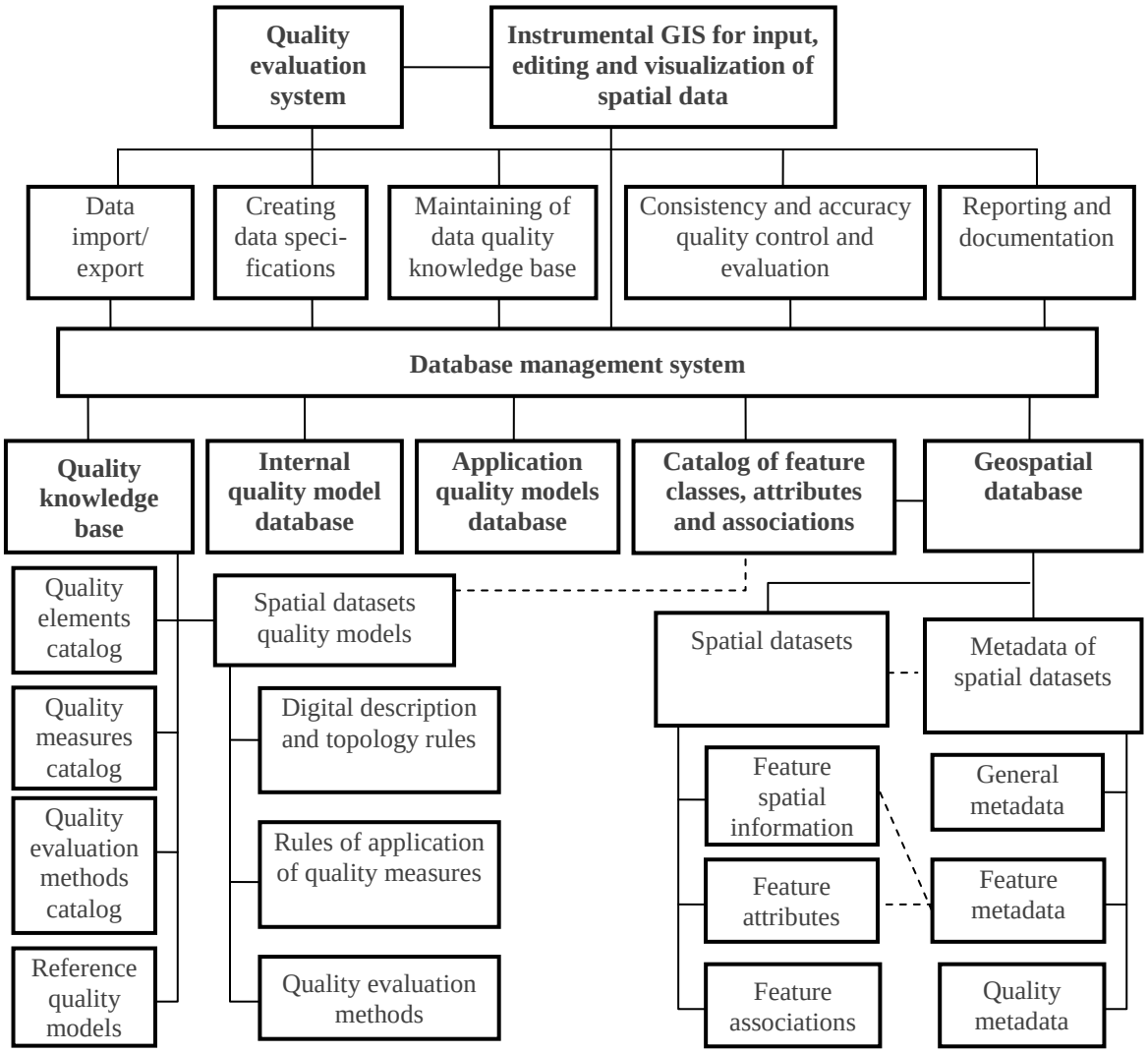

Fig. 2. Functional scheme of automated quality control and evaluation of spatial data

Knowledge base about the quality of spatial data includes the following information resources: catalogs of standardized quality elements and sub-elements, measures and methods for assessing the quality that meet the requirements of ISO 19157, reference data quality models for standard and profile spatial datasets, spatial datasets quality models that are loaded and stored in SDB, consisting databases with rules for digital description of objects and their spatial topology, rules for the application of measures and methods for evaluating the quality elements/sub-elements of a spatial datasets.

We emphasize the importance of the formation and use of an electronic feature catalog database as a key resource for maintaining not only the processes of spatial data quality evaluation, but also for control and assurance of integrity of spatial database, domain and topology consistency, thematic accuracy, as the feature catalog is used for creating interface elements of operating input of spatial object interactively and item-by-item control of domain and topology consistency in process of input their spatial properties (coordinate descriptions) and attributes into the spatial database.

Functional subsystems and information resources of system provide support for the spatial data quality evaluation process by common technological scheme (Fig. 3), for which, depending on the phase of the geospatial data lifecycle, we can identify the following four main options for use: 1) formation model of spatial data quality, 2) evaluation data quality in the production process; 3) certification of finished products; 4) evaluating products for compliance with the requirements of the user.

Creating of spatial data quality model can be used as an independent procedure for the purpose of which is to develop the technical specifications for spatial datasets with justification of the choice of elements and measures, limiting quality values, which are determined at the stage of production data and will be used to assess the quality on the next stages of spatial data lifecycle.

Creating quality model is the first step in technological line of data quality evaluation during spatial dataset certification for establishing the level of correspondence data to claimed quality values, or during the determination of applied quality for assessing the suitability for other use, which is different from the stated

in the technical specification. It is natural, that data quality model based on potential customer's requirements, which assesses the ability to use a spatial data set on the new appointment, is forming during determining the applied quality of spatial data.

The package with the spatial dataset for certification should also have product specification, metadata and data quality report. The conformity assessment using the system is performed as follows:

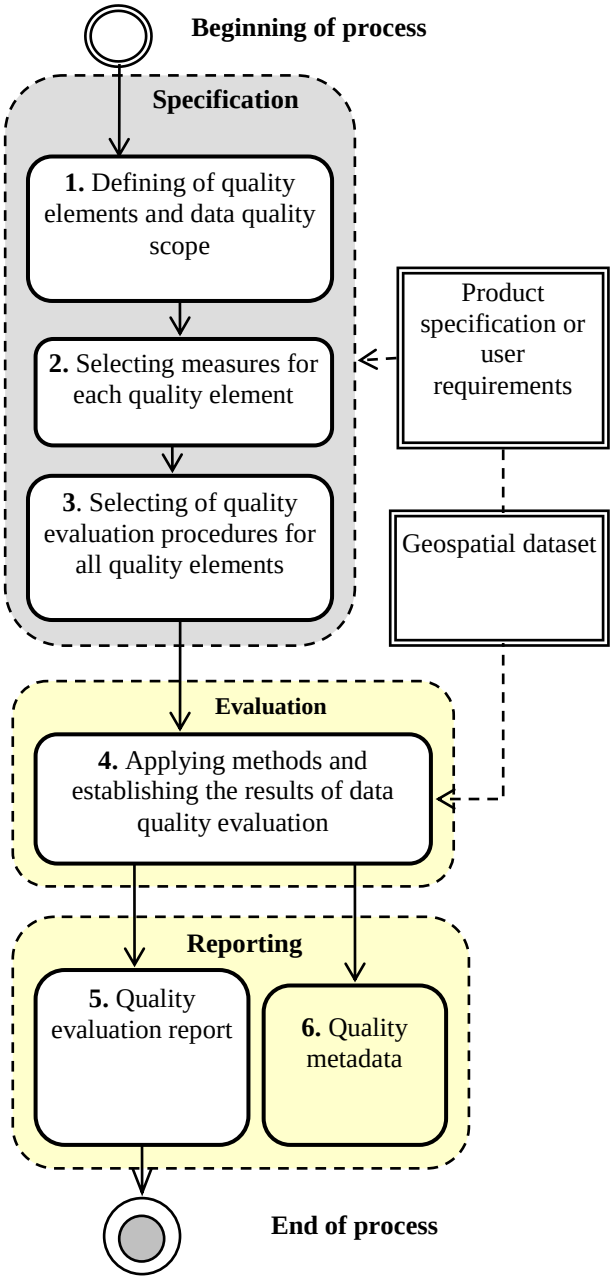

Fig. 3. Overall scheme of geospatial data quality evaluation process

– input of the feature catalog, which is the base for creating dataset, and creating data quality model;

– import and upload of spatial data and metadata;

– quality control of spatial data with detailed report about errors in dataset;

– evaluation of internal quality of spatial data and creating the quality report and metadata;

– comparative analysis of obtained evaluation results and created quality metadata with corresponding quality properties and metadata submitted for certification in a spatial dataset package.

Protocol about conformity/non-conformity of spatial dataset to technical requirements and claimed quality is creating according to the results of comparative analysis.

Conclusions

Geospatial data as modern complex information models of the real world require the creation of adequate automated systems of production, management, monitoring and quality assurance of geographic information products.

Taking into account interdisciplinary nature of the production and use of geospatial data, indicators, measures and methods of quality evaluation should be standardized. The methodological basis of this unification create international standards ISO 9002 series in general questions of quality management and ISO 19100 series in aspects of geographic information products quality.

Creation of automated spatial data quality control and evaluation system on the basis of base-oriented architecture of specialized GIS, which functional model was considered, enables the integration of quality control and evaluation tools into all processes of preparation and production of geospatial data, and provide a high level of independence knowledge based information associated with them and software tools from format and instrumental GIS.

References

1. Karpinskiy J.O. Conceptual principles of spatial data quality evaluation and assurance / J.O. Karpinskiy, A.A. Lyashchenko, M V. Gorkovchuk // Journal Of Geodesy And Cartography. – 2012. – №4. – p. 33 – 41.

2. Lyashchenko A.A. Onthological approach to create the catalog of topographic database / A.A. Lyashchenko, R.M. Runets // Engineering Geodesy – Kyiv: KNUCA, 2008. – № 54. – p. 116 – 123.

3. ISO 19110: 2005(E) Geographic information – Methodology for feature cataloguing. – ISO TC 211, 2005 – 55 p.

4. ISO/DIS 19157: Geographic information – Data quality. – ISO TC 211, 2011.

5. ISO/IEC 13249-3:2002 FDIS. Informational technology – Database languages – SQL Multimedia and Application Packages – Part 3: Spatial, 2 nd edition, 2002.

6. Devillers R., Jeansoulin R. Fundamentals of Spatial Data Quality // ISTE Publishing – 2006 – p. 237 – 253.

7. Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies. EuroGeographics Expert Group on Quality. Edited by Antti Jakobsson, Jørgen Giversen. – EuroGeographics – 2007. – 68 p.

8. Jakobsson, A., Tsoulos, L.. The Role of Quality in Spatial Data Infrastructures. In Proceedings of the 23rd International Cartographic Conference, Moscow, Russia, Cd-Rom. – 2007.

9. Oort, P.A.J. Van. Spatial data quality: from description to application: PhD Thesis. – Wageningen: Wageningen Universiteit. – 2006 – 125 pp.

10. Pro Oracle Spatial for Oracle Database 11g / Ravi Kothuri, Albert Godfrind, and Euro Beinat. – Apres, New York, 2007. – 787 pp. ISBN-13: 978-1-59059-899-3, ISBN-10: 1-59059-899-7.

11. Regina O. Obe and Leo S. Hsu. PostGIS in Action. – Manning Publications, 2011. – 450 p.

12. INSPIRE DS-D2.6, Methodology for the development of data specifications, v3.0, http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/D2.6_v3.0.pdf.

13. ArcGIS Data Reviewer – http://www.esri.com/software/arcgis/extensions/arcgis-data-reviewer/index.html.

Functional model of automated system for geospatial data quality control and evaluation

A. Lyashchenko, M. Gorkovchuk

We consider the structure and function of automated system for geospatial data quality control and evaluation based on the concept of quality geographic information, as defined in the ISO 19100 set of standards, and specialized GIS architecture based on the use of universal RDBMS for storage and processing of geospatial data and knowledge base as a data quality model.