QUALITY BEYOND METADATA- IMPLEMENTING QUALITY IN SPATIAL DATA INSFRASTRUCTURES

Antti Jakobsson[1] antti.jakobsson@eurogeographics.org

Jaana Mäkelä[2] jaana.makela@tkk.fi

Riikka Henriksson[2] riikka.henriksson@tkk.fi

Lysandros Tsoulos[3] lysandro@central.ntua.gr

Matthew Beare[4] Matt.Beare@1Spatial.com

Jorma Marttinen[5] jorma.marttinen@geoaudit.fi

Nicolas Lesage[6] nicolas.lesage@ign.fr

1) EuroGeographics, 6-8 Avenue Blaise Pascal, Champs-sur-Marne, 77455 Marne-la-Vallee, France
2) Helsinki University of Technology, PO.BOX 1200, FI-02015 TKK, Finland
3) National Technical University of Athens, Cartography Laboratory, 9 H.,Polytechniou, 15780 Zographou Campus, Greece
4) 1Spatial, Cavendish House, Cambridge Business Park, Cambridge CB4 0WZ, United Kingdom
5) Geoaudit Oy, Askaistenpolku 2 B 20, 00300 Helsinki, Finland
6) Institute Geographique National, 73, Avenue de Paris, 94165 Saint-Mande, France

Abstract

Currently across Europe, NSDIs are at varying stages of development, data is held in a variety of coordinate reference systems and varies in quality, coverage, content and structure. Quality can be defined as fitness for use, including quality of design, conformance to the design (production oriented quality), customer satisfaction and the fulfilment of the needs of society or environment (Jakobsson,2006). The European Commission's ambition is to build a European Spatial Data Infrastructure (ESDI) on the National Spatial Data Infrastructures in Member States, for which INSPIRE is the legal instrument. This is currently under development and metadata is the first part that should be implemented. Reference datasets are series of datasets that everyone involved with geographic information uses to reference his/her own data as part of their work (FGDC, 2005; Rase et al., 2002). They provide a common link between applications and thereby provide a mechanism for sharing knowledge and information amongst people. INSPIRE will be based on this approach and reference datasets are in a key role in the implementation process. The role of quality in SDIs has been discussed by Jakobsson and Tsoulos (2007) and EuroGeographics Knowledge Exchange Network on Quality has published a guideline on implementing the ISO 19100 quality standards in National Mapping and Cadastral Agencies (Jakobsson, Giversen, 2007).

This paper will explain how a number of European National Mapping and Cadastral Agencies (NMCAs), technology providers and universities will:

- Provide users of reference information with harmonized metadata concepts for discovery and data evaluation purposes.
- Describe a common approach for data quality of reference information at large and small scales. This will be carried out by developing a quality model based on best practices at the data providers in Europe and international standards (ISO 19113, ISO 19114, ISO/TS 19138).
- Develop a quality model with a set of quality measures to be used for reference information at large and small scales. This common set of quality measures is a basis for comparison of quality of reference information between countries and different data providers.
- Develop a web-based semi-automatic evaluator service concept for reference information. This includes logical consistency checking and other quality measures. Web-based semi-automatic evaluator service concept will give guidance for the users of reference information on how data evaluation metadata should be utilized.

The work is part of the ESDIN project (www.esdin.eu) funded by eContent*plus* programme. The project started in 2008 and will end in 2011.

The methodology is based on investigating the user requirements and the best practices at the National Mapping Agencies along with an empirical study of developed concepts. Expected results of the project include a common quality model, guidelines how to set and meet quality requirements in a Spatial Data Infrastructure and how a web-based semi-automatic evaluator concept can be utilized both by data providers and data users. It is expected that quality model and quality requirements will be available in 2009.

This paper elaborates on the semi-automatic quality evaluation process and the way quality model and quality requirements can be used to test quality and report results as metadata.

ESDIN General Quality Model

The purpose of the ESDIN general quality model is to establish a standard approach for describing spatial data quality of reference information at large and small scales. The model helps spatial data producers to harmonize their data quality evaluation processes, the quality measures used and enables reporting quality for INSPIRE Annex I themes. The model is based on best practices at the NMCAs in Europe; on the requirements and recommendations in the INSPIRE Data Specifications and on the international quality standards, such as ISO 19100 series, ISO 2859 and ISO 3951. The quality model defines data quality elements/subelements, data quality measures, data quality evaluation procedures and principles of reporting data quality.

The general quality model consists of the quality model document and quality tables in a spreadsheet format. The quality tables include the following INSPIRE Annex I Themes: geographical names, administrative units, cadastral parcels, transport network and hydrography. Each table includes one theme with all the features and related attributes defined in INSPIRE Data Specifications. Although the current tables include only five of the above-mentioned themes, the structure of the tables is a general-purpose one and can also be adapted for use in other themes.

Quality information should be provided for feature types and attributes for the ISO 19113 data quality elements; logical consistency, completeness, positional accuracy, thematic accuracy and temporal accuracy. Logical consistency is mandatory, the other are voidable.

There are two levels of evaluation metadata that may be provided. The first and the most common case would be a dataset level metadata reporting quality for the feature types and attributes. For completeness, positional accuracy and thematic accuracy quality measures reported are based on ***the conformance levels determined from quality requirements***. The conformance levels are validated by quality tests accomplished by sampling. For logical consistency and temporal accuracy the conformance results or ***the actual evaluation results*** are provided. Figure 2 illustrates the dataset level metadata reported for the whole dataset (scope dataset) and for a subset (scope: sample area).

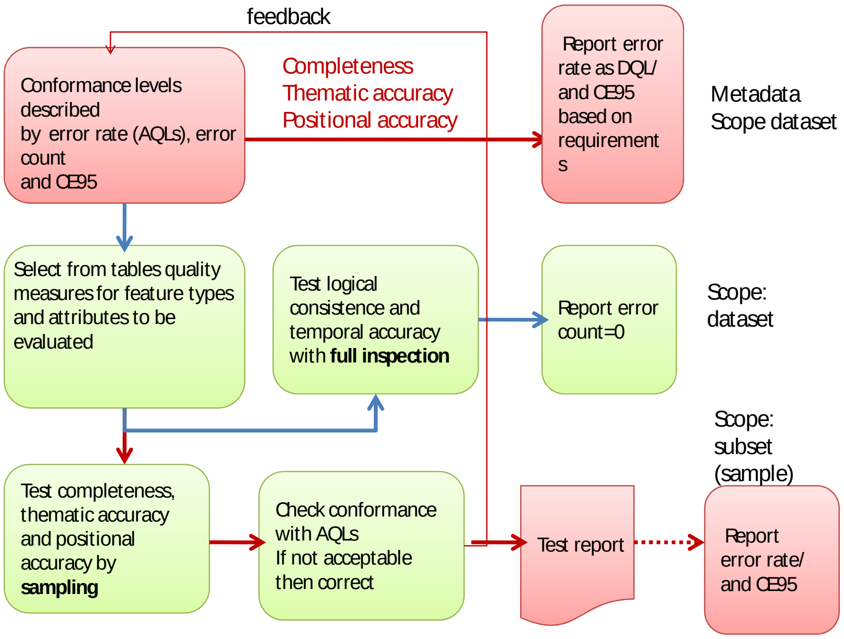
feedback
Conformance levels described by error rate (AQLs), error count and CE95
Completeness
Thematic accuracy
Positional accuracy
Report error rate as DQL/ and CE95 based on requirement s
Metadata
Scope dataset
Select from tables quality measures for feature types and attributes to be evaluated
Test logical consistence and temporal accuracy with full inspection
Report error count=0
Scope: dataset
Scope: subset (sample)
Test completeness, thematic accuracy and positional accuracy by sampling
Check conformance with AQLs If not acceptable then correct
Test report
Report error rate/ and CE95

Figure 1. Dataset level quality evaluation metadata and subset level quality evaluation metadata

The basic quality measures for reporting the conformance results are based on a set of basic quality measures defined in the ISO TS 19138. Those are error rate, error count and CE95. Error rate is used for completeness and thematic accuracy where sampling is used to test the dataset. The error rates reported for the whole dataset should be set using declared quality levels (DQLs).

A producer should report the quality results as conformance levels using the quality measures, which are relevant to the basic quality measures. If no conformance levels have been set, the producer may report subset level metadata that is only applicable for a certain data quality scope (usually defined by an area, but may also be relevant to other scopes).
In practice the producer will report the conformance levels if metadata is provided at a dataset level for completeness, positional accuracy and thematic accuracy.

The basic quality measures for testing are error count, and mean value of positional accuracy. These quality measures are used for testing purposes and changed to error rate, error count and CE95 when reporting.

For the feature types/attributes of each Annex I theme we explain the basic quality measures for reporting purposes.

Setting a Quality Model for a Dataset

The quality model should be designed and formulated before the actual production of spatial data in order to take into account the user requirements and the quality objectives. In essence, the quality model formulates the specifications of quality requirements at entity level, detects the sources of possible faults that affect to the quality of data and specifies the measurements required by the quality assurance operations.

The quality model should include four basic inter-dependent parts:

- **A.** The first part concerns the drafting of the product’s technical specifications or the incorporation of existing technical specifications. The product specifications should include the following:
- The data model (Conceptual model),
- The logical model of the database,
- Rules for filtering spatial information; these rules involve the choice of the real world natural objects to be acquired, so that they are incorporated in the spatial data,
- Rules for the acquisition of spatial information. They involve the methods and processes that are used during the collection of spatial information and that also incorporate restrictions placed by the mechanical equipment, which is used in the collection of elements.

- **B.** The second part refers to the identification of the objectives and requirements of quality at feature type level as these result from the requirements of the product’s specifications. The usual practice follows two distinct steps.

1.Analysis and identification of quality requirements accompanied by the relevant documentation. This is usually achieved with the use of questionnaires or even interviews given by data users.

In order to formulate the quality model using the ISO 19100 series, the selection of the quality parameters requires a (prior) determination of:

- the data quality elements and sub-elements that can be used for the evaluation of the quality of data according to ISO 19113,
- the quality measures that will be applied according to ISO 19138,
- the acceptance criteria and conformity levels of quality. These conformity levels may be set as declared quality levels (DQLs) that are then reported in metadata.

2.When the data quality elements and sub-elements are identified, they are analyzed and compared, with already acquired knowledge, in order to evaluate if they are applicable.

C. The third part addresses the processes for the production of spatial data

The production processes include:

- The drafting and use of an inspection manual (quality control handbook) that describes the quality model to be applied, the processes delineating the model’s application as well as the forms and/or lists for controlling the recording of quality measures and their results (Quality control records - QCR)
- The quality control procedures during the production.

In order for a quality model to be designed correctly, certain basic conditions must be adopted and observed so that the model can de used. Such conditions are:

- For an objective and explicit quality description, each quality requirement should be described with clarity by a quality element and sub element.
- Quality should be described by a constant number of quality sub elements that will depend on the type of data (geometric, thematic, temporal). This practically means that the designed quality model should be applied uniformly/similarly in the totality of data, its categories, its subsets, the feature types, the feature classes as well as its attributes and characteristics (such as relationships).

- The existing models and in general, concepts that have already been formulated and adopted should be taken into consideration.
- The quality model should be applied similarly by all possible institutions.
- The information included in the data refers to the collection time.

D. The fourth part addresses the quality evaluation of the spatial data

- The evaluation procedures according to ISO 19114. Quality evaluation is performed using evaluation methods, in order to determine the quality results using quality measurements either with or without reference data/or reality. This concern the quality measures as these were selected in part B. If reference data is used it should be at least three times more accurate than the data evaluated.
- The production and metadata recording process (ISO 19115 or quality report). Metadata should be incorporated in the production processes and the used software and should include information about the conformity or non-conformity of data with the product specifications.

The evaluation procedure includes:

- Setting the testing schemes based on the ISO standards or some other specifications
- Performing the testing using sampling of full inspection. Here it should be considered using an independent party to perform the tests.
- Monitoring the AQLs and making corrections to the production processes if needed
- Reporting the conformance and/or quality results using the quality measures. Note: if the quality requirements are set, these may be reported as DQLs. The evaluation process confirms that the producer meets these levels. Then the actual results may be kept by the producer, and not reported as a part of the metadata.

Web-based semi-automatic evaluator

The previous chapters explain the principles guiding the implementation of quality evaluation process. The main challenge from producers’ and users’ point of view is how this can be carried out. In the ESDIN project we are working to create a concept for web-based semi/automatic evaluator service that may be utilised by the producer to evaluate the results and support the provision of metadata and by the users to relate producers’ metadata with their requirements. Figure 2 depicts the use case model for the web-based semi-automatic evaluator.

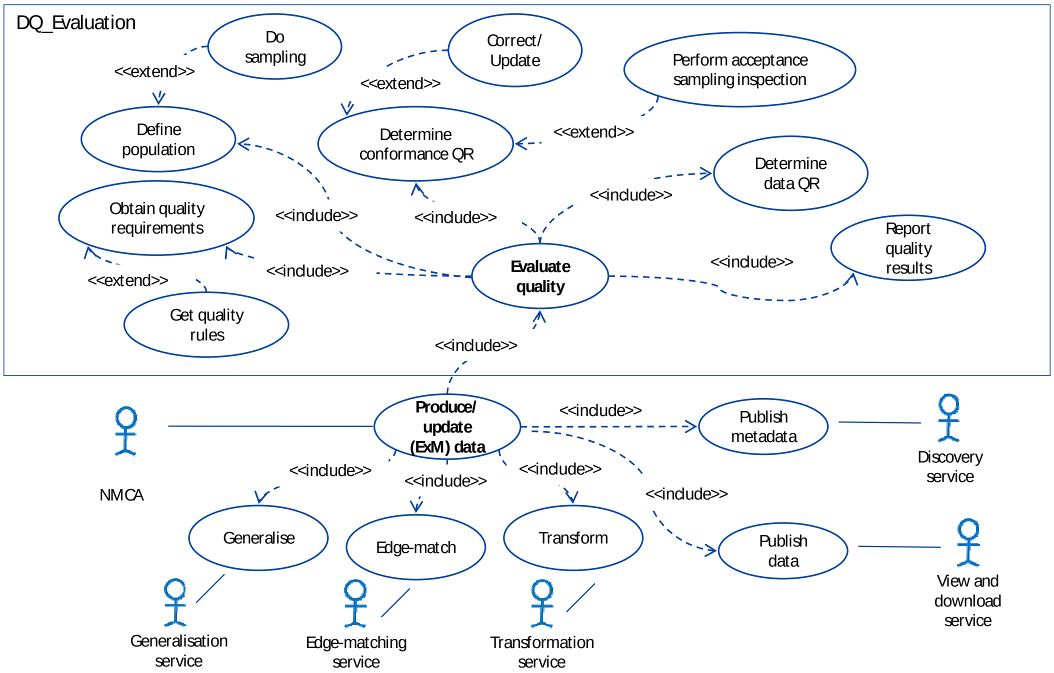

Figure2 Use case model for the web-based semi-automatic evaluator

A web based data quality evaluation service is envisaged to be required in three possible user domains:

- A data provider will wish to supplement the data they are publishing with metadata stating how good the data is, with respect to a product specification and a known set of application purposes (as set out in their quality model);
- A data consumer may wish to evaluate the data available, for its suitability to their particular purpose (which may not have been known by the data provider);
- An auditor may which to evaluate data independently, so as to accredit the data with respect to recognised quality models, data specifications and/or specific application purposes.

The semi-automatic data quality evaluation service is of particular importance to the data provider. An automated service that can be integrated into their data maintenance and data publishing processes has the potential to provide improved operational cost/time efficiencies and increased quality levels. This has the potential to enable faster and more frequent data updates, driving up the currency of data (sometimes argued to be more important than accuracy).

The automatic element of the evaluation service, primarily for logical consistency and completeness, will offer the user an intuitive rules language to express the quality measures and constraints of the data. To support portability, the rules will be expressed in terms of the conceptual data models, without the need for knowledge of or dependency on, the physical models. The service will allow the user to collate and manage sets of rules, and support collaboration across the organisation or user community. The service will automatically evaluate source data with respect to specified rulesets. It should be possible to collate data from different sources, so as to assess data for comparative completeness against reference data. Quality metrics will be published in accordance with ISO 19115 and 19139 metadata standards, to supplement existing metadata. This will allow data users to not only discover what data exists, and where it is, but to also assess how good it is (for their purpose). The evaluation service will be presented through standards based user interfaces and web service interfaces, providing an interoperable service in support of interoperable data.

Quality Metadata

Finally, we will provide guidelines on how quality information should be reported for discovery and evaluation purposes. These will be based on ISO 19115 and INSPIRE Implementing Rules on metadata. We will also study ongoing work in revising ISO 19115 and make an inventory of requirements and practices. The challenge is how to meet different requirements of metadata in the different phases of usage.

Conclusions

A common approach to quality in SDI using a standard approach is considered feasible. This can be considered as a step towards process integration approach discussed in Jakobsson (2006). A common quality model is defined which may be used by producers for the provision of INSPIRE Annex I and other themes. A set of common quality measures are described and then associated to Annex I feature types. This enables users and applications to compare datasets from different producers (sources) and to combine datasets. Part of this work is the utilisation of the results and the development of a web-based semi-automatic evaluator service which may be utilised by both producers and users. Quality information is probably the most valuable part of the metadata but not so much effort has been devoted to the harmonization of quality information. Our approach is a first step towards the harmonization of the European SDIs.

References

FGDC, 2005. Framework Introduction and Guide (Handbook). Digital version, http://www.fgdc.gov/framework/frameworkintroguide/, (accessed July 28th, 2009)

Jakobsson, A., 2006. *On the Future of Topographic Base Information Management in Finland and Europe*, Doctorate thesis, Publications of the National Land Survey of Finland no. 101, http://lib.tkk.fi/Diss/2006/isbn9512282062.pdf, 180p + annexes

Jakobsson A. and J. Giversen eds. 2007. *Guideline for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies*. Eurogeographics Expert Group on Quality.

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

Rase, D., Björnsson A., Probert, M. and M-F. Haupt, eds., 2002. Reference Data and Metadata
Position Paper. Eurostat.