

A COHESIVE APPROACH TOWARDS QUALITY ASSESSMENT OF SPATIAL DATA

JAKOBSSON A.(1), BEAR M.(2), MARTTINEN J.(3), ONSTEIN E.(4), TSOULOS L.(5), WILLIAMS F.(6)
 (1) EuroGeographics, BRUSSELS, BELGIUM ; (2) ISpatial, CAMBRIDGE, UNITED KINGDOM ; (3) GeoAudit, HELSINKI, FINLAND ; (4) Statens Kartverk, HONEFOSS, NORWAY ; (5) Technical University of Athens, ATHENS, GREECE ; (6) Institut Géographique National, SAINT-MANDÉ, FRANCE

1. BACKGROUND AND OBJECTIVES

Since the invention of Geographic Information Systems huge amounts of spatial data have been produced and made available by the public and private sectors. In the early steps of this activity the problem was that there was not sufficient data available to accommodate the wide variety of needs of the users' community. Nowadays the problem is the opposite: Data coverage is almost complete, even at large scales, but nobody can give detailed information on the quality of this data. Provided that quality of a dataset is – amongst others – an indication of its fitness for use, it becomes evident that when this critical factor is missing, one cannot proceed with the use of any data set without risking the results of the application at hand.

It is well understood by all those involved with the application of ISO standards that - although they provide the framework for the standardization of a process and its components - their implementation is a task that requires thorough understanding of a number of factors i.e. the nature of spatial data and their representation in digital environment, technological capabilities relevant to data collection, organization and evaluation, etc. Recent research on this issue has shown that although the application of standards on geographic information is being adopted by the National Mapping and Cadastral Agencies [NMCAs] and the private sector, these organizations are rather far from adopting standards, which refer to spatial data quality. This is due to a number of reasons like: lack of expertise, adherence to internally established practices and - of course – cost.

The work described in this paper elaborates on a cohesive approach towards the quality assessment of spatial data. It was carried out in the framework of the ESDIN project and provides a mature and pragmatic guidance relevant to the problem domain.

2. APPROACH AND METHODS

The overall objective of the ESDIN work on quality has been to provide NMCAs with a paradigm and a comprehensive set of guidelines to assist in handling INSPIRE Annex I themes quality, based on international standards. It is well understood by all those involved with the application of ISO standards that - although they provide the framework for the standardization of a process and its components - their implementation is a task that requires thorough understanding of a number of factors i.e. the nature of spatial data and their representation in digital environment, technological capabilities relevant to data collection, organization and evaluation, etc.

The methodological approach to the problem at hand includes the identification and analysis of the phases required for spatial data quality evaluation as part of a SDI. These phases are the following:

- User requirements study
- Quality model guidelines
- Quality model structure and implementation
- Identification of data quality measures
- Quality measures evaluation methods
- Setting of testing plans
- Setting of quality requirements
- Development of a semi-automatic data quality evaluation service
- Establishing Metadata guidelines

These phases are interdependent in that each one has an influence on several others. This has been a key element of the work done that led to the revision of the quality model and its components until it became “stable” and in accordance with the specifications. Here we discuss further the approach on developing the quality model and addressing user requirements. Identification of data quality measures, evaluation methods, setting of testing plans and quality requirements are discussed further in the report (ESDIN, 2010b). The methodology followed by the project also supports the general framework of quality in SDIs suggested by Jakobsson (2006) and further elaborated by Jakobsson & Tsoulos (2009).

2.1 Developing quality model and metadata guidelines based on user requirement study

Figure 1 presents the main concept on the way quality model guidelines and metadata guidelines are used in the quality evaluation process. It introduces our view of semi-automatic quality evaluation process comprising of quality model (to be described by data producers), data quality web service (for evaluation), reporting results in metadata and finally a data user web service, which would help data users to evaluate quality of data. These two web services are fundamentally different. The first one utilizes ISO 19113, ISO 19114 and ISO 19138 for evaluation of data quality, using quality measures. It includes parts that can be checked automatically like conformance to rules and parts that require manual checking like completeness and positional accuracy testing. The second one uses the results and quality measures for the user evaluation process. This includes understanding of user requirements by setting up a usability model. In the ESDIN project some work was done for the identification of the basic principles of this usability model but still further research is required to have a clear understanding of how usability and data quality concepts can be combined. Eventually, it is considered that a “fit for purpose” service is needed but this was out of scope of ESDIN quality work.

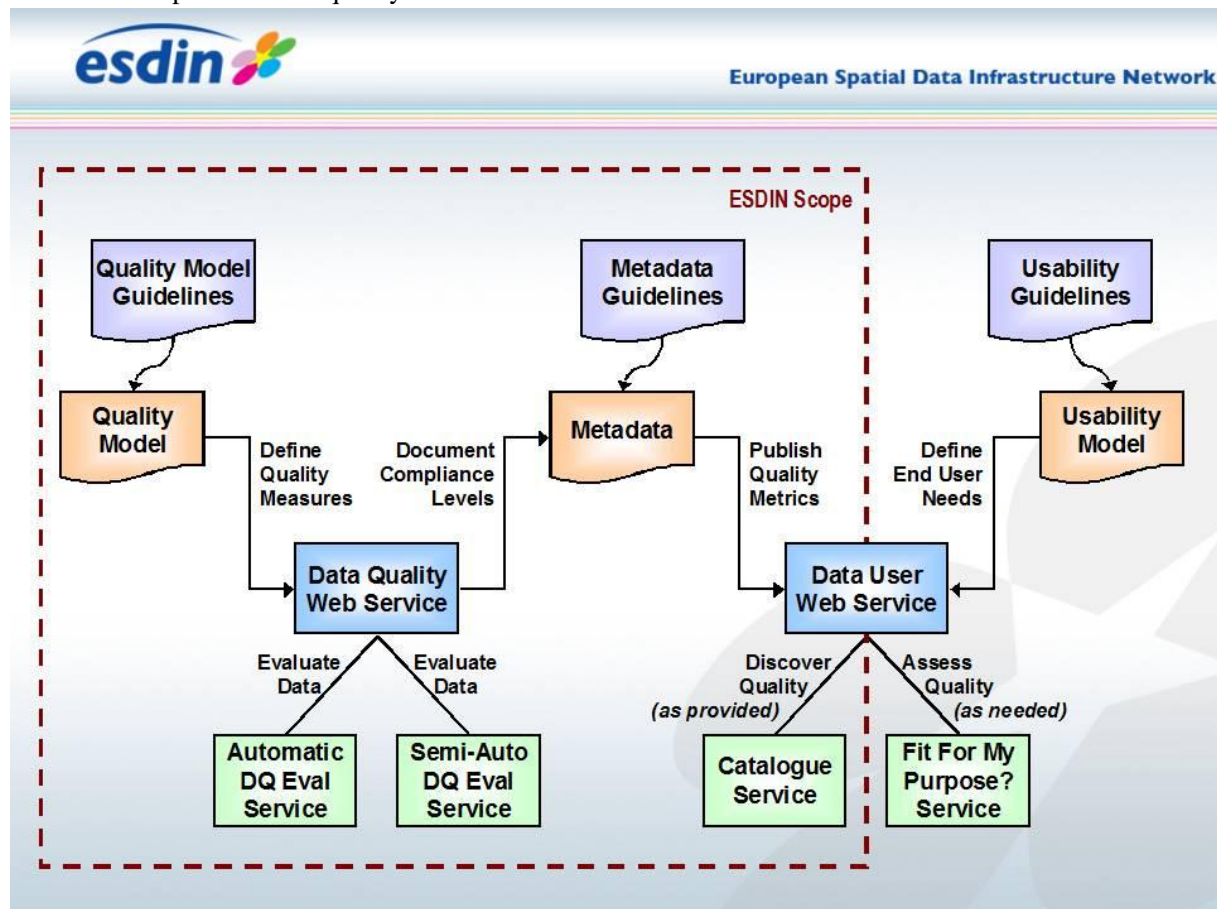


Figure 1 - Concept of quality evaluation in ESDIN

General principles describing a quality model can be developed. The approach has been through setting quality requirements in the quality model by using common quality measures. The quality requirements should be set using users’ requirements studies. Another approach could be just reporting quality evaluation results without setting specific requirements. This has been the INSPIRE approach. While in the ESDIN project the goal is to be INSPIRE compliant, it is our assumption that users want to see some harmonization if they want to utilize data for areas covering many countries or even Europe. The main challenge is to agree on a set of common quality requirements that would satisfy a number of key users.

As it has been already mentioned, the usability model requires more work especially in the area of how users’ needs can be transformed to quality requirements. If only data quality results are reported it may not meet requirements of users who want to use data originating from different sources. Also, if the quality requirements are not demanding we end up with unsatisfied customers. Also, it has to be taken into account that quality levels cannot be changed quickly; a quality improvement project normally takes a long time requiring also major investments from the data providers.

Two different methods of identifying user requirements (and quality measures) were utilized:

- a. Desktop study of existing material from various sources

b. Study of existing specifications at large and small scales

The original idea was to use these studies to identify the quality measures that would be used. When the results were studied it was concluded that the present situation is not mature enough and the original idea had to be revised. Instead of using existing specifications for setting common quality measures the proposal is now based on the ISO 19138 basic quality measures.

2.2 Quality Model structure

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

The quality model should include two basic interdependent parts:

I. The first part refers to the identification of the objectives and requirements of quality at feature type level as they result from the requirements of the product specifications.

The usual practice follows two distinct steps:

i. 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 data quality series, the selection of quality parameters requires the (prior) identification of:

- The data quality 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), which are then reported in metadata.

ii. When the data quality elements are identified, they are analyzed and compared with already acquired knowledge in order to ensure that they are applicable.

II. The second part addresses the evaluation of the spatial data.

- The evaluation procedures are carried out 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 comparison to reference data/or reality. 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).

Quality metadata should be incorporated in the production process and the software used 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 ISO standards or some other specifications;
- Performing the testing using sampling or 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 quality requirements are set, these may be reported as Declared Quality Levels (DQLs). The evaluation process confirms that the data produced comply with these levels. Then the actual results may be kept by the producer and not reported as part of the metadata.

2.3 Development of a semi-automatic Data Quality Evaluation Service

The purpose of the spatial data quality evaluation work within ESDIN has been to develop practical means and recommendations for the introduction of a more automated approach to overcome the risks and challenges e.g.

- measuring quantitative data quality information and reporting quality results as metadata can be a demanding and time consuming task
- quality evaluation still is largely a manual process
- the costs involved in assessing quality
- the existing automation scripts are scripts developed in-house often undocumented
- difficult to provide assurances that the measures applied by an organization are consistent with measures applied by another
- interpretation of standards is difficult

The concepts proposed are the ones that seek to:

- Utilize state of the art technologies and standards;
- Adopt and propose automated techniques;
- Offer shared and consistent evaluation services.

The primary goals sought from this approach are to support the:

- Reuse and standardization of the process;
- Reduction in operational data maintenance and production costs;
- Reduction in product turnaround times and supply chain delivery processes;
- Increased trust and confidence that data consumers have in the available data.

A data quality evaluation service concept is divided into two parts: an automated service and a semi-automated service. This was proposed and illustrated earlier in Figure 1 that shows the data quality evaluation service as a web service capable of evaluating data in terms of the measures defined by the quality model.

Evaluation metrics providing quantitative information on compliance levels exhibited by the data should be documented as metadata in accordance with the metadata guidelines developed. These can then be used by discovery services for users not only to find out what data exists, but also to understand how good that data is and whether it is suitable for their intended use.

The automated service part focuses on those data quality elements and measures that can be fully automated (e.g. logical consistency), and the semi-automated service focuses on those data quality elements and measures that require some human intervention in order to be measured (e.g. positional accuracy).

In the context of a spatial information supply chain, data change is captured and updates to core datasets and products are applied as part of a data maintenance process. For effective data sharing the updated data is typically pushed through any number of transformation and integration processes, each of which may result in the final desired outcome, or provide an input into a subsequent process. This is illustrated in Figure 2.

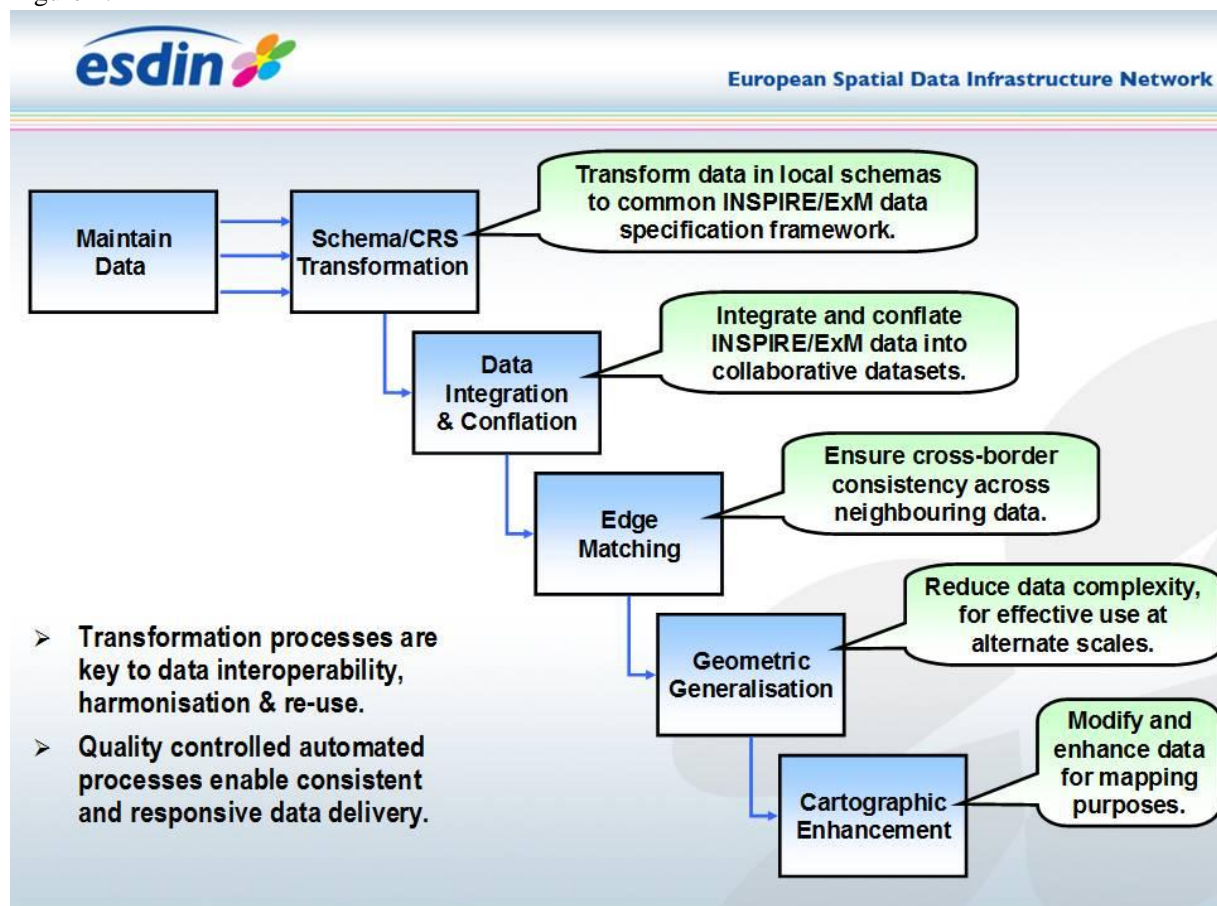


Figure 2 – Transformation processes as part of a data supply chain

These transformation and integration processes are crucial to data sharing initiatives like INSPIRE and ESDIN, which want to re-use available data for a variety of purposes. An important consideration is that as data is transformed, it is essential that its integrity and semantics are preserved.

Quality control is therefore an important consideration for every process that is required to update, modify or transform data in some way. The degree to which the requirements of the process have been fulfilled should be assessed and assurances over the characteristics of the data output can be provided. This observation is consistent with the aims of the ISO standards for quality management systems. The ISO 9000 and 9001 standards define the following relevant terms:

- Product: “Result of a process” (e.g. new or modified spatial dataset).
- Process: “Set of interrelated or interacting activities, which transform inputs into outputs”.
- Quality: “Degree to which a set of inherent characteristics fulfils requirements”.

Given this understanding, data quality evaluation can be considered to be a control mechanism, that provides confidence that any given process has updated, transformed or integrated data in a manner that is expected and that it meets the requirements of the consumer or next process in the chain.

The approach starts by agreeing the quality targets (mission) and defining a set of spatial business rules (quality measures) that can adequately assess the data within the scope of the quality objectives. Guidelines for these first steps are covered in the ESDIN quality model and measure identification is explained earlier.

Subsequently, a baseline assessment analyses the data to determine whether it achieves those targets. If it does not, a cycle is entered where non-conforming errors are reconciled (automatically and/or manually). When the data is corrected sufficiently with respect to the quality targets, the data can be certified as ready for publication.

For subsequent updates rule discovery looks for new rules that can be identified to further improve the quality of data and so the continuous improvement cycle begins.

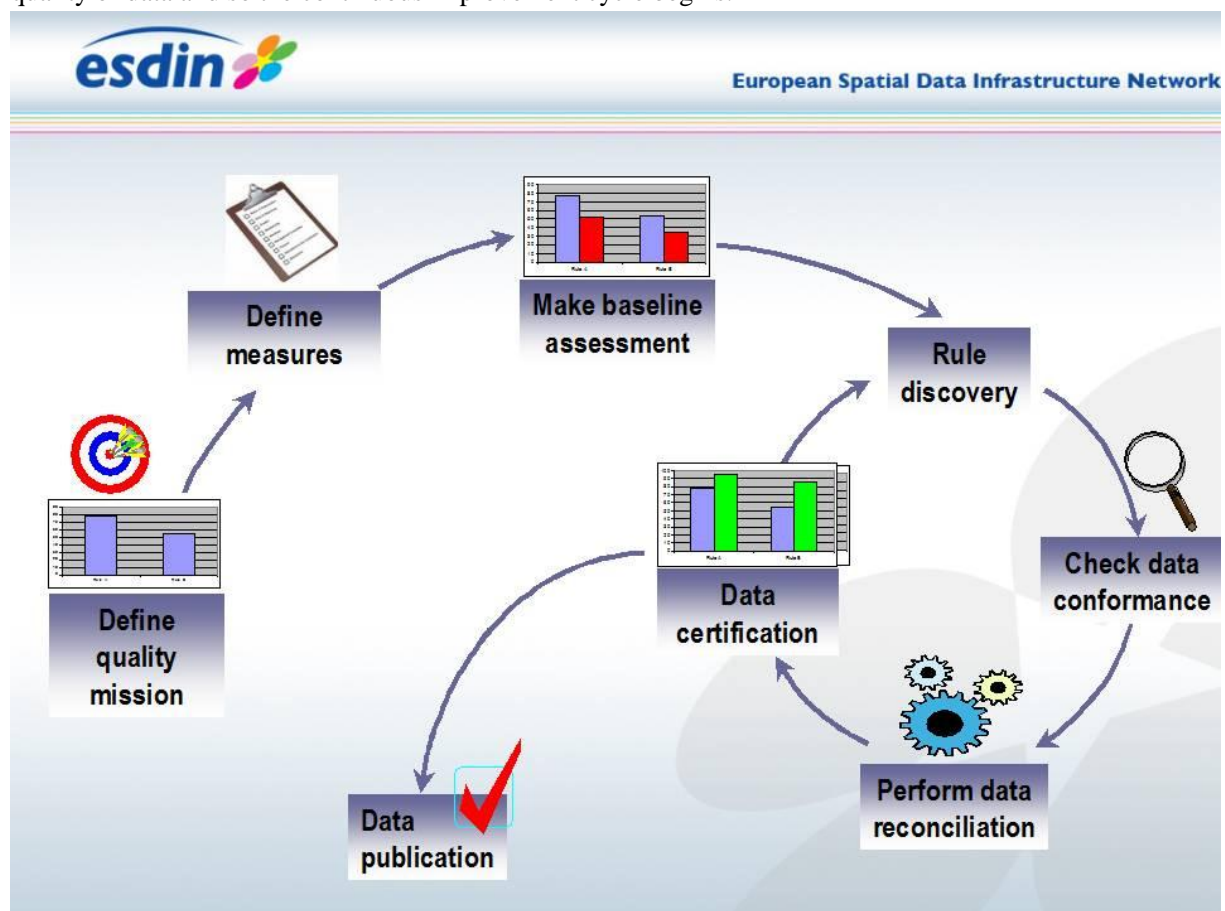


Figure 3 – Data quality improvement cycle

By deploying an effective data quality improvement cycle into the spatial data supply chain and publishing the results, consumers of data can trust and be confident in that data, increasing the likelihood that they will want to use that data and that decisions they make using it will be good ones.

To achieve the best quality, control measures should be applied consistently and often, as data is updated and as it progresses through the supply chain. Frequency and consistency are both difficult to achieve, particularly when doing it manually using different people, or even automatically when using different tools (as is the case across different organizations).

The solution proposed by ESDIN for a consistent cost effective data quality evaluation service is to embrace automation and to provide this automation as readily accessible shared web services.

2.4 Establishing Metadata guidelines

ISO 19115 defines metadata as “data about data”. This basic definition implies an unlimited scope to what can be considered as metadata. Pragmatically, geospatial metadata can be seen as information describing spatial and, to a certain extent, non-spatial resources, making it possible to discover, record and use them.

The aim of the metadata guidelines undertaken by ESDIN was to define a common vision and detailed guidelines on the metadata elements and their contents that should be included with all datasets produced by NMCAs to meet the ESDIN INSPIRE-aligned data specifications.

Furthermore ESDIN is able to provide guidance to end users on the metadata that should be made available for discovery, overview and evaluation of datasets.

The metadata requirements proposed address three conceptual levels of metadata:

- Metadata needed for discovery of datasets through metadata catalogues and registries
- Metadata needed for the evaluation of those datasets, as to whether they are of sufficient quality to meet end users’ needs
- Metadata specific to the requirements of the ESDIN specifications

3 RESULTS

The approach described in this paper is the result of the work of a number of experts in the field from across Europe and provides:

- A cohesive model to enable evaluation and discovery of spatial data based on quality
- Guidelines for setting the quality requirements
- Metadata specification
- A semi-automatic Web based data quality evaluation tool for reference information

In this paper the focus is on quality model and automation of quality evaluation. The guidelines for quality requirements and metadata specification are presented in the ESDIN report (ESDIN, 2010a).

3.1 The ESDIN quality model and test plans

The ESDIN quality model and the service concept constitute a major step in the development of quality management of geographic information.

Figure 4 provides an illustration of the quality model spreadsheet for Geographical Names as an example. A complete set of spreadsheets that support the quality model are provided in the reports. These spreadsheets place the INSPIRE data model elements on the vertical axis and the ISO 19113 Data Quality Elements/Sub-elements on the horizontal axis. The cells of the spreadsheets will then become a “link” between the data model parts and the quality elements/sub-elements. Where the ESDIN Data Quality Guidelines recommend data quality to be measured, the “linking” cells are coloured, and the proper Data Quality Measures are noted, with reference to ISO/TS 19138.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	GEOGRAPHICAL NAMES															
2		DATA QUALITY ELEMENTS														
3		COMPLETENESS			LOGICAL CONSISTENCY			POSITIONAL ACCURACY			TEMPORAL ACCURACY			THEMATIC ACCURACY		
4	FEATURE TYPE & Attributes	COMMISSION	OMISSION	CONCEPTUAL CON	DOMAIN CONSIST	FORMAT CON	TOPOLOG	ABSOLUTE ACCU	RELAT	GROD	ACCUR	TEMP	TEMP	PLA	RE	
5																
6																
7	NamedPlace		DQ basic measure error rate: Id 7	DQ basic measure error count: Id 10												
8	inspireId				DQ basic measure error count: Id 15											
9	name (GeographicalName)															
10	geometry(GM_Object)							DQ basic measure CE95: Id 45								
11	type (NamedPlaceTypeValue)													DQ basic measure error rate: Id 17		
12	localType (LocalisedCharacterString)				DQ basic measure error count: Id 15											
13	RelatedSpatialObject (Identifier)				DQ basic measure error count: Id 15											
14	leastDetailedViewingScale				DQ basic measure error count: Id 15											
15	mostDetailedViewingScale				DQ basic measure error count: Id 15											
16	beginLifespanVersion				DQ basic measure error count: Id 15											
17	endLifespanVersion				DQ basic measure error count: Id 15											
18	Attributes of Data type GeographicalName															
19	spelling (SpellingOfName)															
20	language				DQ basic measure error count: Id 15											
21	nativeness				DQ basic measure error count: Id 15											
22	nameStatus				DQ basic measure error count: Id 15											
23	sourceOfName				DQ basic measure error count: Id 15											
24	pronunciation (PronunciationOfName)															
25	grammaticalGender				DQ basic measure error count: Id 15											
26	grammaticalNumber				DQ basic measure error count: Id 15											
27	SpellingOfName															
28	text															
29	script				DQ basic measure error count: Id 15			DQ hm error count: Id 19						DQ basic measure error rate: Id 17		
30	transliterationSchema				DQ basic measure error count: Id 15											
31	Attributes of Data type PronunciationOfName															
32	phoneticTranscription															
33	phoneticTranscriptionIPA							DQ hm error count: Id 19								

Figure 4 - Example of the Quality Model Spreadsheet for Geographical Names

In order to utilize the common data quality measures, the framework provided by the quality model spread sheets can be used to develop test plans. An example test plan is given based on the INSPIRE Annex I data specification for Cadastral Parcels. The test plans should be set as part of the data specification. ESDIN data specifications give then more comprehensive test plans for large, medium and small scales.

3.2 Semi-automated DQ Evaluation Service

A semi-automated demo service was implemented as a part of the work. The demo employs open source software: the service side is based on the 52North Web Processing Service (WPS) and the client is build upon the OpenJUMP GIS. The items under inspection are x,y coordinates of a point feature (i.e. absolute positional accuracy). The demo service implements three web processes:

- Simple random sampling;
- Measuring circular error at 95 % significance level (CE95);
- ISO 3951 quality conformance evaluation (for testing the AQL levels);

The purpose of the demo was to test the semi-automated quality evaluation concept, particularly the process of ISO 3951 based variable sampling and inspection, in order to be able to determine whether the developed concept works or not and to define the process in more detail and also to demonstrate the idea of semi-automated spatial data quality evaluation.

3.3 Automated DQ Evaluation Service

A fully automated process to evaluate the quality of data as it is processed through a spatial data delivery supply chain has the potential to:

- Generate operational data maintenance and production efficiencies;
- Increase the frequency in which data updates can be delivered;
- Improve the overall quality and currency of available data.

The investment made in developing and deploying such a capability can therefore drive down long term costs and increase customer satisfaction.

Drawing on influences of mainstream IT data management strategies and the adoption of straight through processing techniques, a rule based paradigm can be utilized to drive automated geospatial processing

environments. Within this paradigm knowledge of the data and associated decision logic can be formally expressed as constraints, or rules to which the data is expected to conform. These rules are the quality measures by which the data is to be assessed and through automation can be systematically applied to the data.

By taking the risk of human error out of data processing workflows and allowing computerized services to gather, analyze and process data according to rigorous rules, workloads can be achieved with greater efficiency and more uniform quality. This approach requires a rules language that can effectively express the logical constraints within the domain.

The service requires two interface environments. One to setup and configure the rules; the other to execute the rules. These service environments will have access to the rule repository and the data. This is illustrated in Figure 5.

In the setup and configuration environment, the data quality rules are authored by the domain experts (such as NMCA data analysts/managers). The rules will be derived from the quality measures defined by the data specification team. The environment should be web-based, to support inter-organization collaboration and a sharing of the work load to setup and extend the service rules.

The rules are stored in a data repository. A configuration management process is envisaged, in which candidate rules are authored by the domain experts and then collaboratively agreed and adopted for use by the service. The ability to organize the rules into groups of rules (rule sets) will assist this collaborative review process. It is envisaged that this activity would be overseen by a data custodian with overall responsibility for the quality of data.

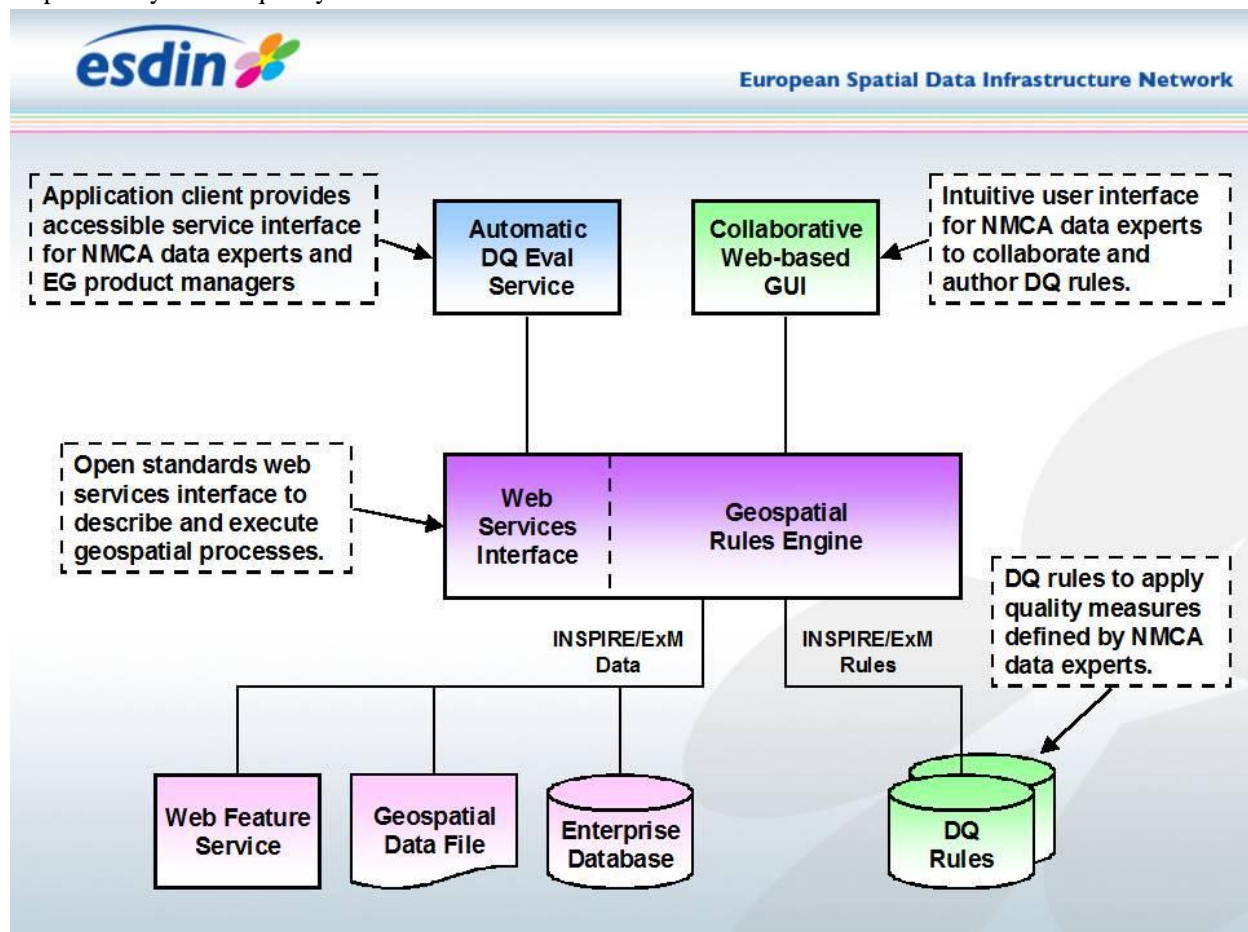


Figure 5 – Components of automatic quality evaluation service

With a set of adopted rules available to the service, the data quality evaluation service can use these to assess the quality of a specified dataset. The service should facilitate the identification of the measures (rules) to apply and the location of the data to be evaluated. In a network service environment, this might be a Web Feature Service (WFS), but additional flexibility to read data from alternative sources may be advantageous, including spatial databases (such as Oracle Spatial or PostGIS) and geospatial data files (such as ESRI SHAPE or MapInfo TAB files).

Over time, it is expected that the rules will need to be extended and enhanced, as part of a continuous improvement programme or as a result of changing data requirements. It is therefore important that the service can readily facilitate this with minimal disruption to the service.

ESDIN project implemented the automated DQ evaluation service with geospatial rules engine offered by Radius Studio. This software was used because it was available to the project and also because it already had implemented the rule language approach needed to implement the quality model.

4 CONCLUSION AND FUTURE PLANS

Utilizing the approach described in this paper the NMCAs can achieve high quality geographic information quickly, advance the profit and advantages of investments and reduce operating costs in the long run. The users will a. rely on data that is trusted and usable b. achieve improved spatial analysis and c. be confident in decision making.

The quality model developed so far covers the INSPIRE Annex I themes. Further work must be done for Annex II and Annex III themes along with a possible refinement of the quality model, the setting of Quality Acceptance Limits for specific applications, the aggregation of quality results and finally full automation of the data quality evaluation service.

REFERENCES

ESDIN, 2010a. D8.3- ESDIN Metadata Guidelines. www.esdin.eu

ESDIN, 2010b. D8.4 ESDIN Quality Final Report – Part A, B, C and D. www.esdin.eu

Jakobsson, A., 2006. On the Future of Topographic Base Information Management in Finland and Europe. Doctoral dissertation. Helsinki University of Technology, <http://lib.tkk.fi/Diss/2006/isbn9512282062/>, 180 p +annexes.

Jakobsson, A., Mäkelä, J., Henriksson, R., Tsoulos, L., Beare, M., Marttinen, J., Lesage, N., 2009. Quality Beyond Metadata – Implementing Quality in Spatial Data Infrastructures. In Proceedings of the 24th International Cartographic Conference, Santiago de Chile, Chile, Cd-Rom.

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.