CO-179

QUALITY OF GEOGRAPHIC INFORMATION - SIMPLE CONCEPT MADE COMPLEX BY THE CONTEXT

TÓTH K., TOMAS R.

European Commisssion, Joint Research Centre, ISPRA, ITALY

1 Introduction

Geographic information is increasingly being shared by many users across different fields and applications. The first milestone is marked by the possibility to combine and overlay different datasets in one spatial environment, while the second by the diffusion of web services that support linking information to geographic location and sharing it with the widest audience.

In this context the reliability of data, which is tightly coupled with its quality, becomes of paramount interest. Morrison (1988), Aronoff (1989), and Longley et al. (1999) identified five main reasons why:

1. Increasing exchange and use of spatial data;

2. Growing group of users that are less aware of spatial data quality;

3. GIS enable using spatial data in all sorts of applications, regardless the original purpose;

4. Current GIS doesn’t offer tools for handling spatial quality;

5. Increasing distance between the end users and those who are best informed about the quality (the producers).

The notion of data quality (DQ) is being transformed – in addition to addressing the a priori requirements of data production, the need for reporting the fitness for (re)use has opened a new approach. The latter is especially pertinent in the context of Spatial Data and Information Infrastructures (SDIs) and volunteered GI collection. One of the objectives of SDIs might be to provide reliable reference data and services to the users to increase quality of the collected information and the associated decisions.

Even though the term “data quality” seems to be trivial, it is rather difficult to discuss because of the assumptions, incoherent terminology, and the diverging viewpoints present in the topic. This paper extends the notion of DQ for SDIs emphasising the similarities and differences with DQ in classical data production. A possible way of dealing with DQ in SDIs will be described using the example of INSPIRE.

2 Data Quality – but which?

2.1 Viewpoints on data quality

Data quality can be described from different viewpoints. Garvin (1988) suggested six (transcendent, manufacturing-based, product-based, value-based, competition-based, and user-based) viewpoints, while Jakobsson (2006), following Lillrank’s idea proposed only four (production-, planning-, customer, - and system-centred). Even though each of the categories can be justified one can argue that sophisticated systems help understanding. However all the categories above can be aggregated in two simple and handy viewpoints: the internal and external ones. The internal viewpoint is related to the data collection and transformation process performed by data producers and providers, while the external viewpoint describes the aspects necessary for reusing the data.

In the classical data production chain producers deliver data that directly fulfils the requirements of a specific user to implement a well-defined task. The a priori requirements are conceptually formalised in the related data product specification, which describes those entities of the real world that are in interest of the original user exactly with the necessary level of details. The selection criteria, the data quality elements with their measures and results as well as the quality assurance system guarantee meting the requirements of the user. Ideally, at the end of the process the results of quality inspections, conformity statement to the data product specification are published as metadata for evaluation.

The classical metadata production benefits of publishing some data specification elements (e.g. title, abstract, etc) as metadata for discovery and the results of quality inspections as metadata for evaluation. This is the point where majority of data producers stops, even though they could go further providing the first input for metadata for use: the description of the use-case or the application that has triggered the data production.

The community of data providers and users interested in data sharing have their first contact through the metadata. As Devillers et al (2007) pointed out non specialist users find difficult to read, understand and map metadata to their requirements. Conformance statements against the original data product specifications do not solve the situation, especially when the potential users do not come from the GI community. In environmental sciences the “scientific” datasets are frequently created without prior specification, never the less they may supply essential information for many users. Since collecting metadata by automated evaluation of data quality is almost impossible, the authors’ (scientists’) accounts on the purpose and the way of data collection are essential.

The new technologies (e.g. GPS, web-mapping, etc) allow spontaneous data collection and sharing, giving way to “volunteered geographic information” Needless to say that the classical data quality documentation does not make part of such data management somewhat hampering further meaningful usage.

Data users need a specific view that helps them to evaluate how much the data fulfil their actual requirements. Since requirements differ from user to user, it is clear that more than one description can be valid. Each of them provides an external view on data and its quality constituting further metadata elements for use.

Opposite to internal data quality, external data quality descriptions do not have long traditions and lack agreed methodology for reporting. The “fitness for use” concept is widely referred, but frequently in wrong context neglecting the difference between the internal and external viewpoints. By exception of describing the original use-case such statements should genuinely come from the users. This data quality element should be documented in a succinct, but informative way. If well-defined user requirements exist they should be appropriately referred and conformity with them should be stated.

2.2 The two faces of data quality: requirements and metadata

When data is produced according to a specification the documentation usually contains parts related to ‘a priori’ requirements fixing strict target results for selected data quality measures to be followed during the production.

Metadata for evaluation and use provides ‘a posteriori’ statements about data quality based on direct measurements, calculations, specific aggregation rules, and other knowledge, expressed as non quantitative information. Metadata includes one or more data quality elements, each of them expressed by a selected data quality measure and the corresponding data quality result. It should be noted that conformance statement is a specific data quality result that is related either to a selected data quality element or to all elements applicable to the inspected dataset.

The connection between a priori data quality requirements and metadata is straightforward. They share the same conceptual basis - the same data quality elements are used both for specifying a priori requirements and reporting data quality as metadata; therefore it is practical to use the same measures for both in the same data set. However this frequently leads to confusion in discussion. This problem is pertinent in spatial data infrastructures, too, where both concepts are related to existing data .

For better understanding the internal and external viewpoints as well as the a priori and a posteriori data quality concepts figure 1 presents the full data production – use – re-use cycle with typical steps and products of the information flow.

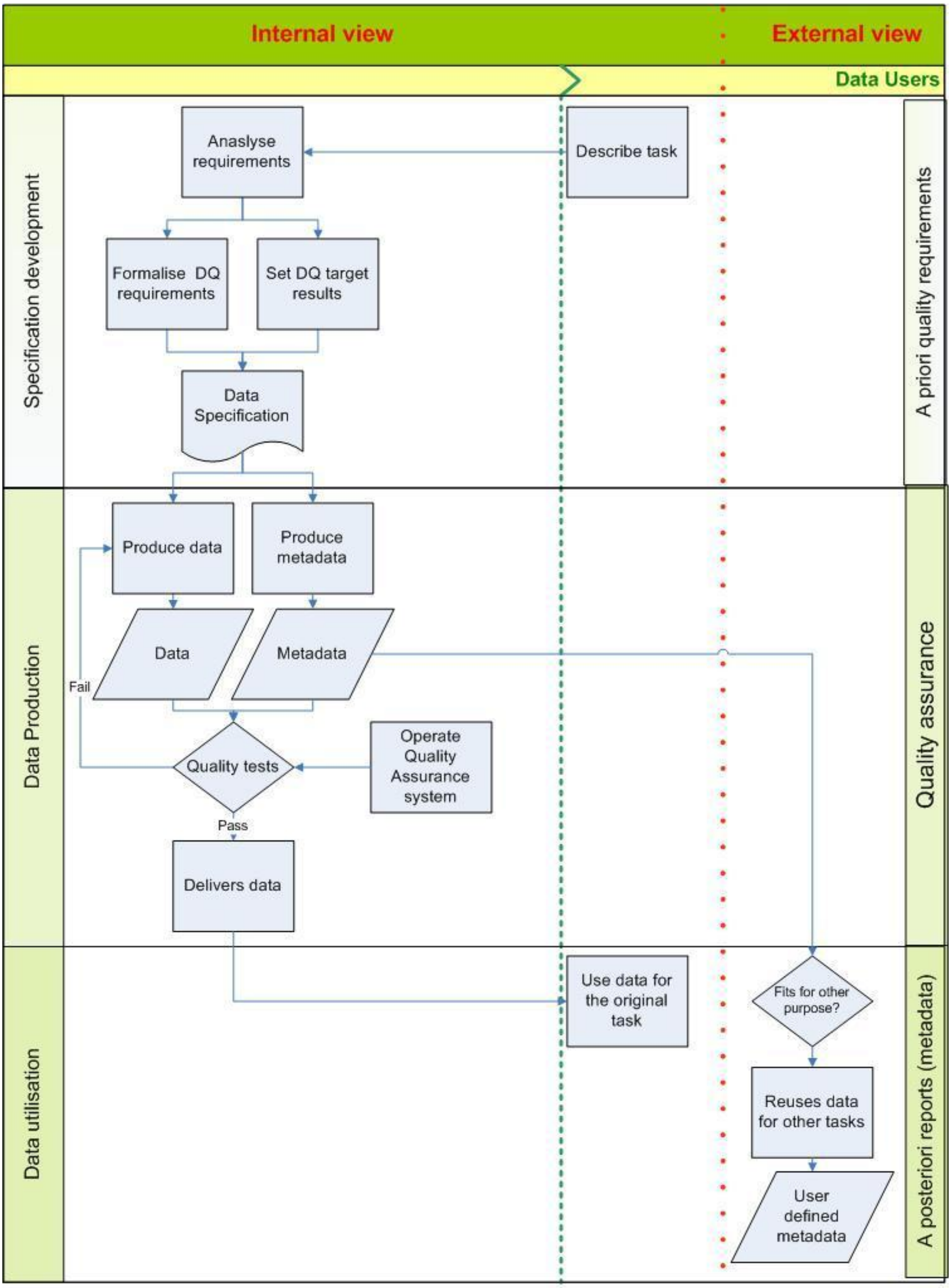


Fig.1. Steps and products in production of geographic information

2.3 Right term, right form

Discussing data quality is further complicated by the lack of commonly accepted terminology. Spatial data and information is being produced by different thematic communities that have developed their own terms. In case of geographic information the relevant existing (ISO 19113, 19115, 19138) and upcoming (19157, 19158) standards of ISO TC 211 provide a good basis for coherent data quality descriptions. However, influenced by national traditions or because of lack of understanding some terms are misused. A typical example of inconsistent terminology is replacing the ISO based “accuracy” by “precision”, ignoring that the first relates to measurements while the second to the properties of the measuring instrument. Another typical example is using “lineage” metadata element as general container of quality statements that do not fit in the data quality elements defined by the ISO standards.

The problem of inconsistent terminology is further aggravated when data quality statements come from different user communities. For example, one would expect the Earth observation community to use similar descriptors to those used in GI. Even though there is no fundamental difference in the approach, the Quality Assurance Framework for Earth Observation of GEO (QA4EO) promotes using “quality indictors” based on general standards of metrology instead of the ISO TC211 family.

Some data providers consider that detailed metadata can be replaced by conformity statements, but it should be noted that that conformity at data set level is a “compound” metadata element which reflects adherence to the cited specification. Internal conformance statements are useful to expert users that know the contents of the referred specifications and have the necessary knowledge to understand and interpret the underlying data quality measures and their results.

Non-specialist users may benefit from conformance statements against user requirements, or standalone quality evaluation reports. A conformance statement against user requirements can be defined as external conformity. This latter has not been widely used so far mainly because of the lack of formalisation and mechanisms for collecting such elements. The proposed “DQ_DescriptiveResult” of the upcoming ISO 19157 allows evaluation in a descriptive text without forcing the quantitative ways. Although it is a “subjective” evaluation by nature, it is easy to read and does not require formal testing or strict evaluation procedures.

The textual metadata descriptions are increasingly recommended by different authors and communities. It should be noted that the “fitness for purpose” and “usability” concepts are vague and difficult to quantify. Standard structures of content in a form of templates and report layouts would help the users to quickly orient themselves, which should be underpinned by such interfaces on the geoportals that allows querying text elements in a distributed environment.

3 Quality in spatial data infrastructures

An SDI provides the technical and legal framework for accessing and reusing spatial data. It is assumed that it is built on existing data originating from disparate sources. Ideally an SDI provides access to data in interoperable way, i.e. without the need for specific ad-hoc interaction of humans or machines. The interoperability target is formalised in interoperability (data) specifications that follow the structure of the data product specifications.

As compared to data production the role of ‘a priori’ data quality requirements in SDI is different. When establishing the data component of an SDI, two aspects need to be balanced:

1. Giving access to the widest selection of data;

2. Achieving interoperability.

The second condition comprises data quality requirements too. The transformations necessary to achieve the targeted data structure should not compromise the original quality of data in general.

A priori data quality requirements may play a discriminative role when deciding whether to include a specific data set in an SDI. Balancing the wide spread publication of data with requirements against the quality is a delicate decision that directly influences the content of specifications for interoperability. Consequently data quality requirements may be completely absent from the interoperability specifications when the main purpose of the infrastructure is to share every existing data set.

Contrary to data quality requirements, metadata on data quality is an essential content of every SDI . Nevertheless providing metadata in SDI is not a simple exercise of reprinting the metadata of original data sets. First of all the transformations necessary to fulfil the interoperability may lead to deterioration of original quality. Secondly, there is no one to one mapping between the data sets of the data providers and the “data themes” of the infrastructure. In order to provide data according to the interoperability target several datasets from different sources have to be integrated. As a consequence, the data quality becomes inhomogeneous, which requires dedicated methods and rules for metadata generation.

Updating the original metadata elements might be cumbersome. In topographic data production, for example, it can be based on calculations/aggregations, or on quality inspection based on appropriate sampling . A practical alternative can be using original metadata with information on the process step describing the transformation methods and the possible associated errors, expressed in the MD_Lineage element or in a standalone report.

Interoperability also requires agreed way for measuring and reporting data quality; otherwise it is not possible to compare the metadata associated to different datasets. Therefore the metadata part of the interoperability target specification has to fix the data quality elements, evaluation methods, and a list of measures. If possible, these aspects should be harmonised across the data themes too.

Whatever detailed the metadata descriptions of data quality are they may have a "fairly limited impact on user's ability to understand the possible uses of data," which means that the gap between what the quality assessment experts can produce and the information that the users can understand and use persists. In the current ISO 19115 standard there is an extension to MD_Information called MD_Usage, which allows data providers to describe the usage. This option has not been used by the SDI community so far, but it has the potential to fill the previously mentioned gap.

The concept of usability has been proposed as a new specific DQ element in the upcoming ISO 19157 standard. The new DQ_Usability is defined as a "degree of adherence to a specific set of data quality requirements", or "adherence to a particular application" is especially useful when other data quality elements do not sufficiently address a component of quality. This metadata element allows expressing data quality both in terms of conformance and descriptive results. This is in line with the Quality Assurance Framework for Earth Observation of GEO (QA4EO) that promotes assessment to an agreed reference or measurement standard allowing numeric or text descriptors for presentation.

4 Data quality in INSPIRE

4.1 Reiterative approach

The thematic scope of the INSPIRE Directive is defined in the 34 themes listed in the annexes of the legal act. In December of 2009 interoperability specifications for Annex I have been agreed by the Member States of the European Union.

Combining spatial data from different themes and sources in a consistent way represents a strong data quality demand that go beyond the logical consistency within a data theme. The consequent application of the data modelling elements and other provisions of the Generic Conceptual Model (GCM) enforces cross theme consistency. The GCM lists data quality among the data harmonisation elements, but does not give a generic data quality model and other details for specifying data quality requirements and metadata descriptions.

Based on the general provisions of the GCM the interoperability specifications have been developed by separate expert groups for each team, which was followed by a cross theme harmonisation phase. The topic of data quality requirements proved to be rather difficult because of the need to balance the requirements stemming from the use-cases with the legal boundary conditions stating that INSPIRE has to be built on existing data. As consequence, the expert groups were reluctant to propose a priori data quality requirements even for those specification elements where a clear demand stemming from other EU legal acts existed.

Also mixing the a priori data quality requirements and metadata parts of the data specifications has caused some trouble leading to re-occurring discussions not only in the technical work, but also in various phases of the legal process. In order to guide the discussion a data quality expert group from the representatives of the Member States has been set up. The initial position of the Member States has been clarified by answers to a mini-survey, which also clarified the context and the terminology to be used. After a couple of iterations the final report with recommendations on data quality is expected to be delivered in spring of 2011 providing input for the final phase of the INSPIRE data specification process. The particularities of INSPIRE and the interim results of the work of the data quality expert group are described in the next chapters.

4.2 DQ in the legal acts and the data specifications

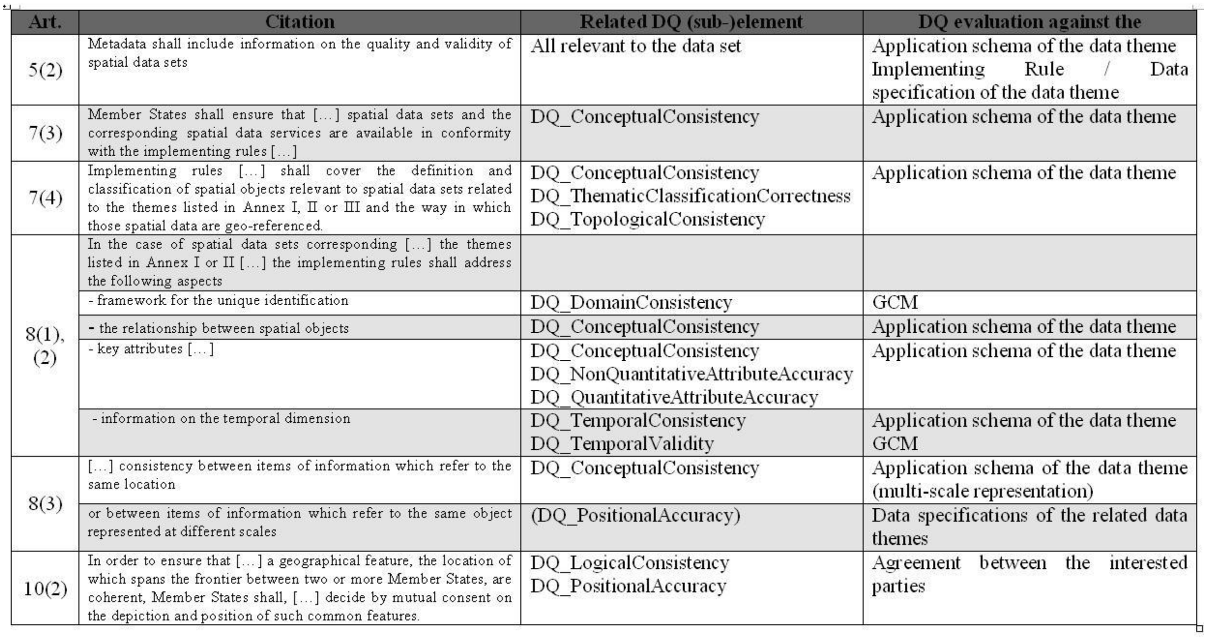
INSPIRE, like any Spatial Data Infrastructure, targets to use data from different providers by multiple users and applications. The Directive contains explicit and implicit requirements related to data quality . Table 1 establishes mapping between the DQ related statements of the legal act and the DQ elements and sub-elements used in the interoperability specifications. Following the "use of existing" principles these data quality elements follow the ISO TC 211standards.

The majority of DQ requirements of the legal act relates to conceptual consistency, ensuring semantic interoperability between the systems of different data providers. Clear definitions of spatial object and data types that are consistent across thematic domains, independent from the data provider, and stable in time constitute a major step towards better usability of data.

All the DQ requirements of INSPIRE has to be reported as metadata. The Conformity specified in Commission Regulation (EC) 1205/2008 as regards on Metadata is reported at dataset level. Since the multiplicity of this metadata element is 1..*, it is possible to report conformity with different specifications. However, neither the metadata, nor the data interoperability regulations contain rules for the aggregation process and miss to link conformity with the appropriate (data quality) specification elements.

For evaluating conformity not only the elements, but also the applicable DQ measures and the related results and tolerance values have to be specified. Guided by the “non exclusion” principle the expert groups responsible for Annex I data specifications have selected DQ measures from ISO 19138, but as a rule have not established targets for their values. The exceptions to this practice are those datasets, where the content is regulated by legal acts; i.e. Administrative Units or Cadastral parcels with their zero tolerance for DQ_Omission.

The missing DQ results constitute a situation when all the data provided in INSPIRE should be 100% correct to be conformant with the Directive. Needless to say that this contradicts to the uncertainty associated with the measurements and classifications present in the GI technology. This requires further work, which may result in refining the DQ and MD chapters of the INSPIRE interoperability specifications.



Art.	Citation	Related DQ (sub-)element	DQ evaluation against the
5(2)	Metadata shall include information on the quality and validity of spatial data sets	All relevant to the data set	Application schema of the data theme Implementing Rule / Data specification of the data theme
7(3)	Member States shall ensure that [...] spatial data sets and the corresponding spatial data services are available in conformity with the implementing rules [...]	DQ_ConceptualConsistency	Application schema of the data theme
7(4)	Implementing rules [...] shall cover the definition and classification of spatial objects relevant to spatial data sets related to the themes listed in Annex I, II or III and the way in which those spatial data are geo-referenced.	DQ_ConceptualConsistency DQ_ThematicClassificationCorrectness DQ_TopologicalConsistency	Application schema of the data theme
8(1), (2)	In the case of spatial data sets corresponding [...] the themes listed in Annex I or II [...] the implementing rules shall address the following aspects		
	- framework for the unique identification	DQ_DomainConsistency	GCM
	- the relationship between spatial objects	DQ_ConceptualConsistency	Application schema of the data theme
	- key attributes [...]	DQ_ConceptualConsistency DQ_NonQuantitativeAttributeAccuracy DQ_QuantitativeAttributeAccuracy	Application schema of the data theme
	- information on the temporal dimension	DQ_TemporalConsistency DQ_TemporalValidity	Application schema of the data theme GCM
8(3)	[...] consistency between items of information which refer to the same location	DQ_ConceptualConsistency	Application schema of the data theme (multi-scale representation)
	or between items of information which refer to the same object represented at different scales	(DQ_PositionalAccuracy)	Data specifications of the related data themes
10(2)	In order to ensure that [...] a geographical feature, the location of which spans the frontier between two or more Member States, are coherent, Member States shall, [...] decide by mutual consent on the depiction and position of such common features.	DQ_LogicalConsistency DQ_PositionalAccuracy	Agreement between the interested parties

Table 1: Data quality statements of the Directive and the related DQ elements, sub-elements, and baselines for evaluation

5 Lessons learnt from INSPIRE

The dialogue with the Member States outlines some preliminary conclusions about the role of data quality in INSPIRE that might be useful for other SDIs too.

The terminology used for describing data quality has to be harmonised. Whenever possible, the terms have to be taken from the related ISO TC211 family. When new terms, especially data quality elements and measures have to be introduced other authentic sources (standards on metrology, terms defined by international organisations) have to be used. The data quality terms have to be introduced in the INSPIRE Glossary and have to be published for the wide audience.

In order to establish a common approach in INSPIRE the GCM has to be complemented, if possible, with a quality model with appropriate references to the related ISO standards.

A priori data quality requirements and recommendations have to be based on representative use-cases and user requirements. DQ requirements have to be set with caution; they may exclude data from the infrastructure. Requirements should stem directly from the Directive, other binding European legal acts, or very strong / well-accepted user requirements. Logical groupings of DQ requirements or recommendations can constitute basis for (user-defined) conformance classes.

A priori requirements on data quality and related MD should be more stringent for reference data than those on thematic data themes. This is justified by the commonly used object referencing that plays important role in referencing thematic data and achieving overall consistency in INSPIRE.

The potential of conformity should be better understood and exploited. Beyond the mandatory conformance class stemming from the legal obligations other classes can be specified with more ambitious targets. When a data specification element is very significant from the point of usability conformity with that element should be reported separately.

Conformance testing should be based on objective criteria that are traceable, quantified, and linked to specific DQ elements. The target results for a priori data quality requirements (DQ_QuantitativeResult) have to be accompanied, when appropriate, with values for tolerances and the description of the selected evaluation method. The assembly of the necessary steps in conformance testing has to be included in the abstract test suits.

The number of metadata elements for evaluation and use should be kept reasonable. Only the most representative metadata elements shall be picked and they should be reported in an easy to understand way.

The transformations necessary to reach interoperability within the infrastructure may lead to deterioration of data quality, which requires updating metadata about the data quality. This can happen with full inspection, sampling, or indirect evaluation. Since the first two methods may require substantial investments, indirect evaluation using original metadata and description of the process steps applied in transformation may be used.

Lineage should not be misused. According to ISO 19115 lineage is information about the events or source data used in constructing the data specified by the scope. Even tough data quality is an aggregate of lineage and DQ elements, lineage cannot be the container of data quality elements that are not specified elsewhere in the data specification. Lineage should be limited to the lineage source and process step.

It is highly desirable to structure descriptive reports. The eventual theme specific templates both for lineage and usability (MD_Usage) should be harmonised across the data themes whenever possible. This may help user’s orientation and machine translation.

Formalising standalone DQ reports provides essential value for the users in deciding whether the selected datasets fit their purpose. DQ_Usability of ISO/CD 19157 follows this idea but misses identify the list of applications where the data has been used, which would give a baseline for comparison and initial orientation.

6 References

Aronoff, S. (1989) Geographic information systems: a management perspective: WDL Publications, Ottawa

COMMISSION REGULATION (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R1205:EN:NOT

Devillers et al (2002) Spatial Data Quality: from Metadata to Quality indicators and

Contextual End-user Manual. OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management, 21-22 March 2002, Istanbul.

http://www.cassini.univ-mrs.fr/publis/OEEPE_ISPRS_Devillers. pdf

Devillers et al (2007) Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. International Journal of Geographical Information Science Vol. 21, No. 3, March 2007, 261–282

GEO Task DA-09-01 Data Management Subtask a.: GEOSS Quality Assurance Strategy http://www.grouponearthobservations.org/cdb/geoss_imp.php, http://qa4eo.org/

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML

INSPIRE Generic Conceptual Model

http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/D2.5_v3_3.pdf

INSPIRE Methodology for the Development of Data Specifications

http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/D2.6_v3.0.pdf

INSPIRE Data Specifications – Guidelines (Annex I)

http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2

ISO/TC211 (2003) 19114 Geographic information – Quality evaluation procedures

ISO/TC211 (2005) 19109 Geographic information – Rules for application schema.

ISO/TC211 (2003) ISO 19115 Geographic information – Metadata

ISO/TC211 (2006) ISO 19139 Geographic information – Metadata – XML schema implementation

ISO/TC211 (2010) CD 19157 Geographic Information – Data quality

Jakobsson, A (2009) Is there quality in SDIs? Presentation at the 11th GSDI Conference, Rotterdam. http://www.gsdi.org/gsdiconf/gsdi11/wrkshpslides/w3.7b.pdf

Jakobsson, A et al (2009). Quality beyond Metadata – Implementing Quality in Spatial Data Infrastructures. In proceedings of the XXIII International Cartographic Conference. Santiago de Chile. http://acreditacion.fisa.cl/icc/contenidos/nonrefereed/3/f_2009363N2PZX1.doc

Longley, P.A.et al (1999). Geographical Information Systems: Principles, Techniques, Applications and Management. Wiley New York

Morrison J.L.(1988) The proposed standard for digital cartographic data. The American Cartographer (15) pp 129-135

Oort, P.v. (2005): Spatial data quality: from description to application. In Publications on Geodesy 60. Nederlandse Commissie voor Geodesie Netherlands Geodetic Commission, Delft

Sanderson, M et al (2009) SDI Communities: Data quality and knowledge sharing. In proceeding of the 11th GSDI conference, Rotterdam http://www.gsdi.org/gsdiconf/gsdi11/papers/pdf/283.pdf