Expanded Data Quality Model for Increased Reliability in Mashed-Up Environments

Jürg H. LÜTHY Switzerland

Key words: Digital Cadastre, GSDI, Data Quality, Service Quality; traceability; integrity; Web-services

SUMMARY

Along with the growing availability of Spatial Data Infrastructures the publication of spatial data through the Open Government Data initiatives has been significantly increased. Thanks to standardised services (Web Map Services, Web Feature Services) and a range of supporting software tools and libraries the consumption of spatial data is uncomplicated.

To ensure correctness of data and correct use of the data at the end user the quality model for the SDI of the Canton of Zurich (Switzerland) must be completely revised and expanded. The quality model covers the roles of data producers, data providers and data users as well as the transport processes in between them (complete data chain). For the producer side it is based on existing concepts (ISO 19157) but is expanded for better addressing the third and fourth dimension. To ensure the data integrity from data origination to the end user the aspect of traceability must be moved from a descriptive metadata item to a core quality element. The real-time use, the spread and the retransmission of data sets requires the introduction of the criterion quality of service which must be addressed by service providers and service brokers.

Governmental and moreover cadastral data must put highest priority for correct, reliable and trustable data and the contingency of data services also under adverse conditions. The quality model presented in this paper serves as guideline for the further development of the SDI of the canton of Zurich.

Expanded Data Quality Model for Increased Reliability in Mashed-Up Environments

Jürg H. LÜTHY Switzerland

1. INTRODUCTION

The aspect of quality in the context of spatial data has been discussed for more than 20 years. In 1988 the National Committee for Digital Cartographic Data Standards published a first standard for digital cartographic data. Since then various standardization organisations and authors published proposals for the description of quality parameters for spatial data (cf Lüthy 2008, ISO 19157:2013).

Due to the shift in use of spatial data – from single sourced digital maps over Spatial Data Infrastructures (SDI) to mash-ups on a mobile device – the quality model focused on the data set itself is no longer adequate in such an environment (Lüthy et al. 2015). Along with the growing availability of SDI the publication of spatial data through the Open Government Data initiatives has been significantly increased. Thanks to standardised services (Web Map Services, Web Feature Services) and a range of supporting software tools and libraries the consumption of spatial data is uncomplicated. But in many cases the consumer of data is neither aware of the data source and its quality nor of the service provider and its service reliability. This tendency is amplified with growing data distribution through mash-ups and integration services. Data sets may be used therefore for purposes for which they are not made for and wrong decisions may be taken. On the other side, data providers becoming also more and more decoupled from the consumer side not knowing how their data is being used. An expansion of the quality model for the modern use of spatial data is therefore appropriate.

1.1 About the term quality

The term quality plays a central role in this paper which requires a common understanding. In common speech the term is used non-judgmental to describe the behaviour of a product or a sevices and evaluating the class of a product. In ISO 9000 the term "quality" is defined as the degree to which a set of inherent characteristics fulfils the requirements (ISO 9000:2015). The description of the quality of a product is therefore tightly coupled with the purpose (fitness for use). Before the quality of a product can be assessed it is necessary to elaborate the requirements and specifications. This tight link can be seen in the ISO Standard Data Product Specification (ISO 19131:2007).

1.2 Data chain

The need for an expanded quality model can also be explained by comparing the data chain according to ISO 19157 and nowadays use of data. In Figure 1 the relatively short and simple data chain in the traditional approach is given. Moreover only a limited number of actors are involved. One of the key elements for such a data set is the common understanding of the abstraction process when transforming real world objects into database features to meet the user requirement. Deviations in data models or feature capturing might lead to errors in data usage, especially when no or only limited metadata are published together with the data. In general it can be seen that the data provision and data usage are in a close context so that semantic differences and misinterpretation are rare.

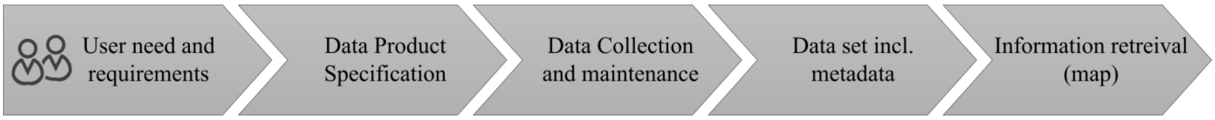
User need and requirements
Data Product Specification
Data Collection and maintenance
Data set incl. metadata
Information retreival (map)

Figure 1 Data chain according to ISO 19157

For the expanded quality model a more complex data chain is taken as a basis. The quality model covers the roles of data producers, data providers and data users as well as the transformation and transport processes in between them. In general the chain between user need and data set is the same. In the SDI data sets are often not published in total but are tailored according to user's needs (step filtering and portrayal in Figure 2). Mistakes made in this process are not yet covered in a model and furthermore the black-boxing of such tasks makes the quality control difficult.
The technical provision of WebServies is often with other organisations than the data team as different domain knowledge is required. The concepts for quality of service as required by European Commission (2007) suits the technical aspects of the provisioning (process quality) but does not address the result aspects (data quality).
Further downstream a service broker may bundles streams from different service providers to provide a meshed-up information package which is supposed to fulfil a customer requirement at the consumer side. Hence a service broker is acting as a kind of data originator. Such combination can theoretically be nested several times but in practical terms the performance of such architecture degrades quickly. Yet, the environment in which the data is used may be completely different to the original purpose of the data set.
The end user has eventually access to a multi-sourced and multi-processed information bundle about which nowadays very little quality information is provided. The missing availability of metadata is probably one of the reasons why users often trust blindly the map displayed on the smart phone. Since metadata are not easy to read and understand and because the users are often disgusted by the terms of use it should be considered that the data provider monitors the appropriate usage of its data sets.

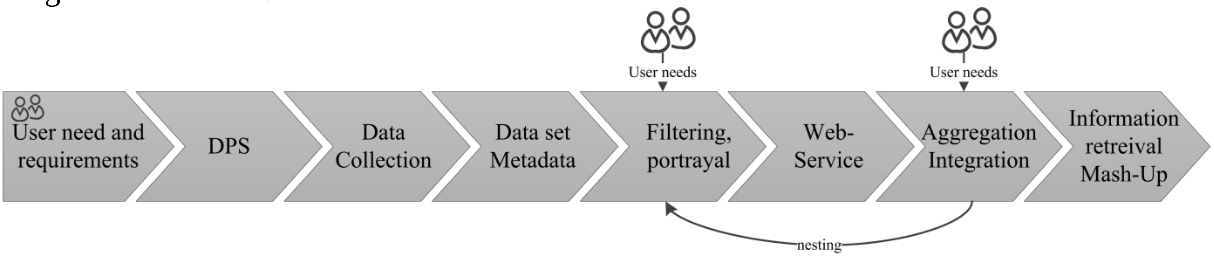

Figure 2 Complex data chain in todays service-oriented enviroment

2. QUALITY MODEL IN FOR A MASHED-UP ENVIRONMENT

To ensure correctness of data and correct use of the data at the end user the quality model for the SDI of the Canton of Zurich (Switzerland) must be completely revised. The developed quality model is built up on the main elements in the data chain as introduced in the previous chapter (see also Figure 3). It covers the roles of data producers, data providers and data users as well as the transport processes in between them (complete data chain). To join the links tighter the overarching quality elements traceability and integrity will be used in each part. The criterion traceability requires that for each object the rational and the history of a feature must be documented. For a data driven community the entire data chain, from origination to the end use must be better controlled. Traceability of information is in domains with high safety requirements like aviation an importance topic (Eurocontrol 2007). Traceability must be achieved on a data level (i.e. supporting documentation for a feature) but also on a configuration level (i.e. documentation of data model, feature capture rules, presentation model). Integrity means that data item (spatial, textual and relation information) is not lost or altered since the data origination or authorised amendment. The amendment can be a persistent change of value or a temporary presentation of a data item through filtering, aggregation and portrayal.

In the following sections the component of the quality model will be described in more details.

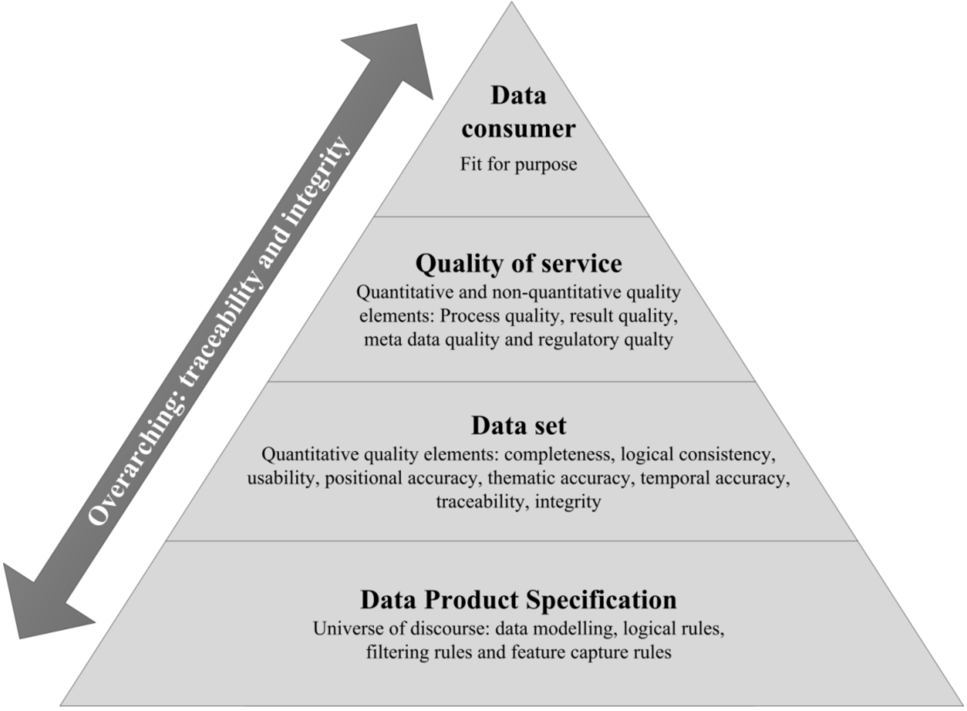
Overarching: traceability and integrity
Data consumer
Fit for purpose
Quality of service
Quantitative and non-quantitative quality elements: Process quality, result quality, meta data quality and regulatory qualty
Data set
Quantitative quality elements: completeness, logical consistency, usability, positional accuracy, thematic accuracy, temporal accuracy, traceability, integrity
Data Product Specification
Universe of discourse: data modelling, logical rules, filtering rules and feature capture rules

Figure 3 Quality Model related to the main data processes

2.1 Data Product Specification

The basis of the quality model is the Data Product Specification (DPS) where the representation of the real world in the spatial data set is defined. The tasks of deriving these requirements are described below. Since the DPS are the starting point for all further activities it is considered indispensable to provide a complete, consistent, comprehensible and unambiguous set of specifications. As the DPS are often related to historical maps or data sets they cannot encompass the entirety of today's usage. The DPS can ensure that the gathered and maintained information fulfil the users requirement for the particular data set.

The core structure and content of the DPS can be taken from ISO 19131. This standard covers all relevant aspects for describing a data set are. With respect to the use of the specified data set in an open environment the data modelling seems to be of primary importance.

Data modelling is the process in which real world objects are analysed by their importance for a *particular* data set (or application) and the dependencies between the objects. Data modelling is often mistaken for the process where a logical database model is derived. Data modelling is much more than that. As the name implies, modelling has something to do with simplification. A (data) model is therefore an abstraction of real world features. Because of limited resources for data processing, data storage and because of the associated costs for data collection and maintenance, the goal of data modelling is to filter those real world properties which are deemed to be necessary. The relevance of a property strongly depends only on its intended use. In the abstraction process the following tasks can be distinguished:

- Data modelling. The "thematic" abstraction defines which properties of a real world feature are important to know and how they are related. Whenever possible domain ranges or enumerations should be part of the database model.
  Sample: For an obstacle database the location, the height and the presence of marking of a TV tower are relevant information for the aviation domain. The fact that there is a restaurant and an observation platform has no importance for an obstacle database (but for a tourist guide).
- Logical rules. In addition to the database model further logical rules for the guarantee of data consistency and integrity such as reference systems, data quality, data maintenance, delivery format and metadata catalogue must be defined.
  Sample : In Inspire the coordinate reference system for compound 3D coordinates shall be expressed in ETRS89 on GRS80 (2D) and European Vertical Reference System (height)
- Filtering rules. The "filtering" abstraction sets the minimum conditions which an object has to fulfil in order to be regarded as a feature worth capturing in the data product. The filtering rules also contain the requested area of coverage.
  Sample: only areas bigger than 1.000 m2 are considered to be a forest. Smaller wooded areas are ignored.
- Feature capture rules. The "mapping" abstraction finally defines how the properties of an object (spatial and non-spatial) must be mapped into the data model.
  Sample: The TV tower might be captured in 2D as a point, as a point with a diameter or as a circular surface. The height could be captured as top and base elevation or as 3D geometry.

The presented abstraction processes result in holistic data product specifications (or requirements). The degree of abstraction and the tolerance for simplification have a big impact not only on the costs for data capturing and maintenance but also on the usability of the data set. Where thresholds are too high, a data set might become unusable for a certain application. If no traceable rules for the abstraction process are available or if different data providers use different rules, an end user has no possibility to find out whether a data set matches his demands or not. Since the elaboration of the DPS is done before the data collection, integrity is not yet of major relevance. But as the DPS may contain specifications for presentations it is beneficial for the long-term use when such rules are defined at an early stage. They can serve downstream as binding specifications for the creation of a web service (see section 2.3.2).

2.2 Quality of a spatial data set

The elements used for specifying and documenting the quality of a data set are based on the ISO 19157 standard. To better account the 3rd and 4th dimension in spatial data the relevance of the elements temporal accuracy and logical consistency will grow and are expanded therefore. As additional elements traceability and integrity are included on a data item and data set level as well. As the content of ISO19157 should be known only the deviations are discussed here.

Logical consistency is the degree of adherence to logical rules of data structure, attribution and relationships. For 3D data sets the topological rules will be much more comprehensive compared to 2D geometry. The definitions and rules for a 3D topology are based on Brugman (Brugman et al. 2011). Also for the correct representation in the fourth dimension the logical consistency plays a central role: for each point in time a valid topological state must be defined. In most cases a state attribute (planned, in use, out of service, dismantled) must be given for each feature in order to support the entire life cycle. A temporality model where all features are temporal with start of life and end of life but where also every features change over time should be considered for effective validation of the topological consistency requirement. Having such attributes for a feature defined it is obvious that the temporal accuracy, i.e. the accuracy of the temporal attributes and temporal relationships of features becomes important. For some data sets it must also be considered that not only the timestamp but also the temporal reference must be given for each change. Where a temporality model is implemented the traceability over the life span of an object can be easier achieved.

At this stage of the process chain the integrity of a data item should be uncomplicated since databases will provide it. Where data exchange is needed in the data collection (survey sensor to database or external data originators) it is suggested to transfer it in an electronic format and to protect it against loss or alteration by the application of a data integrity protection like a cyclic redundancy check (CRC).

The benefit of comprehensive quality requirements can only be fully achieved if acceptable conformance quality levels and appropriate test methods have been defined. Instead of “1 m vertical accuracy” it is better to define “95 % of all values shall have a vertical difference of not more than 1 m compared to ground control points”.

To comply with the quality requirements appropriate processes have to be set up. They are based on feature capture rules but should also address the quality requirement in the corresponding process step. With the practical implementation it turns sometimes out that some assumptions and expectations made in the specification phase cannot be fulfilled: the notification process may is not so good to achieve the completeness requirement. It is recommended to review the DPS and the implementation processes after a trial period to ensure that the data set complies with the DPS.

2.3 Quality of service

As the outcome of the two basic principles a spatial data sets according to DPS and technical requirements is available for use. Data can be consumed from the same organisation which is responsible for the data collection. The broad demand for digital data sets comes often from outside users which satisfied by a SDI through Web services like Web Map Services (WMS) and Web Feature Services (WFS) respectively. The availability of a particular data set, its product specification as well as its quality level can also be provided via the Internet, more specifically via a Catalogue Services for Web (CSW) relying on the Meta Data catalogue according to ISO 19115. The provision of data through Web services requires the introduction of the element service quality. Although the importance of geographic Web services is well-established, their quality is often questionable. Moreover, the issue of geospatial data quality has an important role for the adoption of certain geographic Web services (Medeiros, 2009). There are several proposals for the definition of the quality of service (QoS) like W3C standards organisation (W3C, 2003) or the Inspire team (European Commission 2007). Both approaches are focussed on data delivery infrastructure but less on the data consumer and not addressing the separation between service provider with a detailed knowledge about the data sets and the data user with potentially little knowledge. Hence the QoS must be expanded compared to the Inspire directive. The quality factor model as propose by Wu (Wu et al. 2011) is a promising approach because its hierarchy and flexibility.

The proposed QoS model is built up on the following four factors (see Figure 4): process quality (technical basis), result quality (delivered data), metadata quality (correct description of the data content) and regulatory quality (fitness for use). Service quality must be addressed by service providers and service brokers.

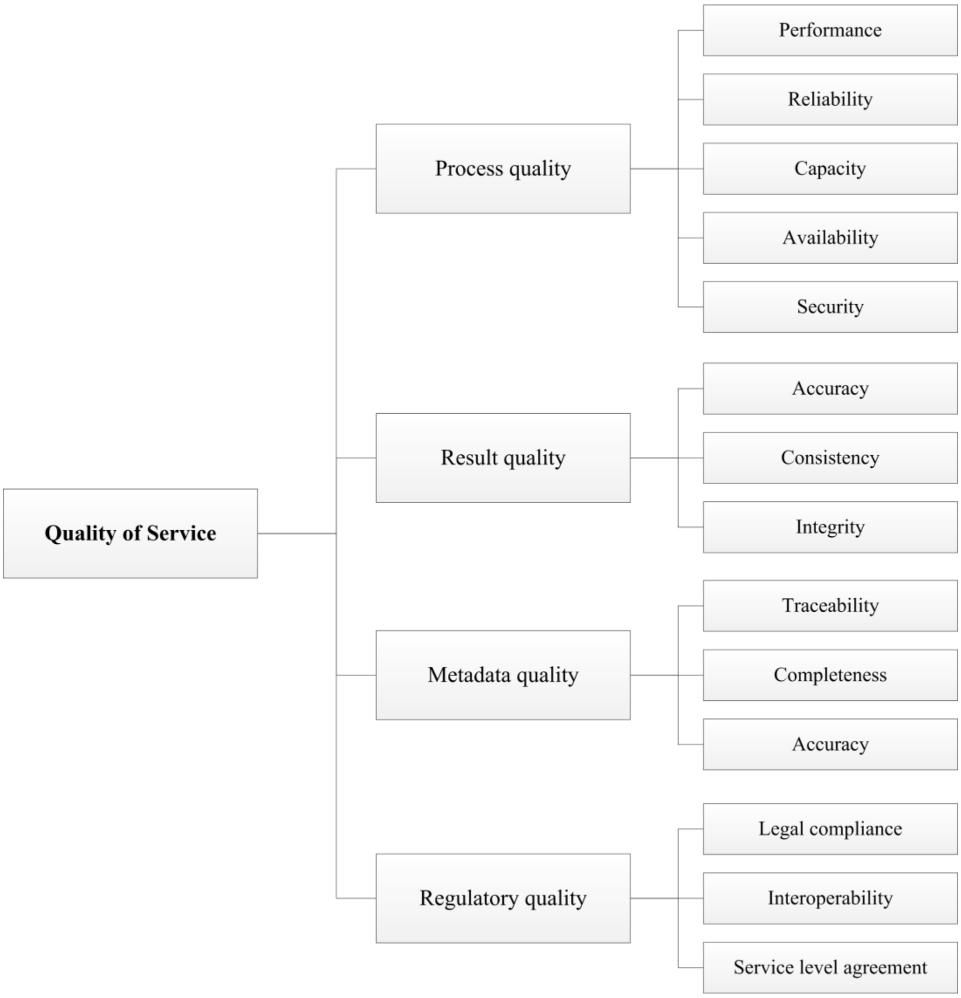

Figure 4 Structure of quality of service

2.3.1 Process quality

The performance of a web service is dependent on the combination if infrastructure at the provider side (server, data structure), user side (client, bandwidth) and in between them (network). None of the actors can measure the overall performance of the system because it is affected by the slowest component (“bottleneck”). Within a SDI the performance of the server and network up to the internet access point can be used as a measurable set of requirements. The perceived quality of the service at the user side should be deduced from these requirements in order to give the end user an approximate performance level.

Therefore, the quality elements for the process quality are defined as follows (see also European Commission 2007):

- Performance: represents how fast a specified service request can be completed;
- Reliability: represents the ability of a web service to perform its required functions under stated conditions for a specified time interval. The reliability is the overall measure of a web service to maintain its service quality;
- Capacity: is the limit of the number of simultaneous requests which should be provided with guaranteed performance;
- Availability: is the probability that the system is up;
- Security: Web services should be provided with the required security, providing confidentiality and non-repudiation by authenticating the parties involved, encrypting messages, and providing access control.

The security element requires access control and authentication of each single user. Most of the SDI are open to the public which means that any user should with very limited technical barriers get access to data. These contradictory requirements cannot be solved at a general level. The Canton of Zurich follows an open government data strategy (OGD) which requires that data open to the public should not be protected by authentication. Within the OGD strategy also cadastral data, rights, restrictions and responsibility will be published. Such data is sensitive because there are financial implications associated with it. Where no authentication and encrypting are used to protect such information data items may be corrupted on its transport to a user through criminal elements. The above-mentioned CRC algorithm can also be used to allow an end user to control the integrity of the provided information. Separate channels for the publication of public uncritical data, public sensitive data and non-public data (for local administrations) must be built up to address the different security requirements of the data.

2.3.2 Result quality

Besides the technical level the provided content has to be correct. In most cases the information published through the SDI are not identical with the originated data set. The data is tailored to different user needs which requires operations on the data like filtering, aggregation and for map services also rendering. According to the concept of a spatially enabled society (Kaufmann et al. 2012) spatial and non-spatial data may be linked together at this stage. Such linking or integration of several sources into one object is beneficial for the end user since holistic information can be provided from one hand. On the other side the process must be handled careful because of potential deviating granularities and because of dependencies to sources outside of the organisation.

Due to the heterogeneous user community there are typically no specific requirements available if not already contained in the original DPS. However it is recommended compiling specifications for the different products so that the correctness of a result can be validated. Following elements can be used to describe the result quality:

- Accuracy: is the quality of the algorithms and operations used to derive the information form the source, the integration of several sources for one object and the degree of adherence to the specifications.

- Consistency: A service may be registered in several registries that support different standards and reference models. Transformation between these different reference models may generate inconsistency.
- Integrity: A user must have the possibility to validate the correctness of a transmitted object to ensure that it has not been corrupted during transfer.

2.3.3 Metadata quality

The need for metadata varies largely between the different user and applications. For a reliable use of the provided information i.e. when decisions are taken based on external information, it is inevitable to publish metadata about the services. Private users (citizen) are usually not aware of the technical aspects and set higher priority on the availability of an information than on the accuracy of it.

- Accuracy: The metadata must accurately reflect the content of the data set. Metadata must be complete and up to date.
- Traceability: To support the data integrity from data origination to the end user the traceability information must be made available. Some of the lineage information must be provided at the item level, other at the data set / web service level. Because of the massive amount of information which may be necessary for traceability this information may only be made available upon request and are not completely contained in the metadata.
- Consistency: The metadata of the service provider is in a registry. This registered information may not coincide with the actual service. Alternatively, the service is updated while the metadata is not updated.

2.3.4 Regulatory quality

Some of the web services of a SDI serve to fulfil a legal mandate. In such cases several requirements may be described besides the abovementioned process and result quality requirements. The regulatory quality requirements are the aspects describing conformance with the rules and the law. They ensure that the technical requirements regarding the interoperability are met like the kind of web service (WMS, WFS etc.) provided and which standard version (WFS 1.1 or WFS 2.0) are supported. The service level agreement (SLA) may also be regarded as a regulatory requirement. The SLA usually contains specifications regarding process quality requirements as described above.

2.4 Data consumer

The simple access to spatial data, the comprehensive standardisation (both on technical level and data content) and the broad selection of tools for querying and visualising spatial data the users often ignores the original purpose for which a data set is collected. The consumers assume that when data are available that they are “correct”. The more the data can be represented as a simple map or meaningful visualisation the less it is considered to validate and verify the appropriateness of a data source for a specific usage. The quality of the information according to the previous sections is – unless further operations are applied – unchanged at the user’s side, but to ensure that the information is “fit for purpose” some specifications should be compiled against which the data source can be compared (validated).

Hence, the data consumers must be sensitised for appropriate selection and combination of data sets. Because with the provision of feature based web-services the control over data is with the user they must be made aware of potential misuse and eventually they must ensure algorithmic accuracy and usability.

How can data provider help building up the awareness? Since requests on web-services can be tracked the content and frequency of data deliveries to an IP-address can be monitored even for anonymous users. If thresholds are exceeded the user request may be deviated to a service site where guidelines, background information and metadata are provided.

3. APPLICATION

Free of charge, non-governmental maps services disclaim the responsibility for continuity of the data service and the correctness of the provided data in their service policies. Governmental and moreover cadastral data must put highest priority for correct, reliable and trustable data and the contingency of data services also under adverse conditions. The quality model presented in this paper serves as guideline for the further development of the SDI of the canton of Zurich and of a regional SDI. For cadastral data the formal aspects of the presented quality model is for large part implemented (DPS including feature capture rules, detailed data model and portrayal rules, continuous quality assessment and corrective action). Following the requirement imposed by Federal Act on Geo-Information all spatial data sets provided by the canton or its municipalities must be technically harmonised. The cantonal authority initiated a project for the development of DPS for each data set. For each topic a project team is put together in which all relevant know how must be represented. Thanks to this approach it can be ensured that the approximate 200 different data sets from all kind of topics (infrastructure, planning, agriculture, biodiversity etc.) will have a harmonised basis. This basis will also include the required quality level for the data sets. Due to the discussion of the different users represented in the project teams a comprehensive view on the required QoS of the cantonal SDI can be built up. On the other side every participant of the project teams will become sensitised for potentials of mash-ups but also on the potential misinterpretation of data if used in the wrong context.

In the next phase a transition from the historically grown SDI to a more specifications and requirements based infrastructure must be initiated. The introduction of the Cadastre 2014 for the entire area of the Canton is an important driver for the implementation of the core quality elements traceability and integrity. This will require fundamental change of existing infrastructure and will therefore not be established for the entire SDI from begin on. But when data sets with formal DPS will be published through the SDI more attention has to be paid on the fulfilment of the QoS factors result quality and regulatory quality.

4. CONCLUSION

Considerations to the quality of spatial data are not new. Changes of technical methods for data storing, capturing, presentation and use require a different view on the quality model. The broad use of spatial data through web-services, the rising importance of mesh-up, data linking and the shift from data providers to service brokers and data users is not covered by existing quality models. The data and service provider have to establish a framework for a fully digital information data supply chain for enabling a modern and reliable cadastral information service. In a future SDI an end user must have the possibility to verify the suitability of the data source for the given purpose and the correctness of data based on a comprehensive documentation of the data chain. Traceability for every data item is for data sets with a high relevance indispensable for ensuring the reliability. The second major improvement for reliability comes with the integrity requirement: The integrity focuses on the process and service quality ensuring that data is not lost, uncontrolled altered or corrupted between origination and consumption. The presented model is a viable approach and in parts already transferred into practise.

REFERENCES

Brugman, B., Tijssen, T., & van Oosterom, P. (2011). Validating a 3D topological structure of a 3D space partition. In Advancing geoinformation science for a changing world (pp. 359-378). Springer Berlin Heidelberg.

Eurocontrol (2007), Final Report for the Draft implementing rule on aeronautical data and information quality, Edition 2.0, 16 October 2007, 237 p. – http://www.eurocontrol.int/sites/default/files/article/content/ documents/single-sky/mandates/20071016-adq-final-report-v2.pdf (last accessed February 10 2016)

European Commission (2007). Inspire network services performance guidelines. INSPIRE Consolidation Team, 2007.

ISO 9000 (2015). ISO 9000:2015, Quality Management. ISO International Organization for Standardization, Geneva.

ISO 19131 (2007). ISO 19131:2007, Geographic information -- Data Product Specifications. ISO International Organization for Standardization, Geneva.

ISO 19157 (2013). ISO 19157:2013, Geographic information -- Data quality. ISO International Organization for Standardization, Geneva.

Kaufmann J. and Steudler D. (2012). Common Data Integration Concept. In Steudler D. and Rajabifard A. (Editors). Spatially Enabled Society, FIG Report 58, FIG Denmark, Copenhagen, pp 23-28.

Lüthy, H. J.. (2008). Entwicklung eines Qualitätsmodells für die Generierung von Digitalen Geländemodellen aus Airborne Laser Scanning (Vol. Nr. 95, Mitteilungen / Institut für Geodäsie und Photogrammetrie an der Eidgenössischen Technischen Hochschule Zürich). Zürich.

Lüthy H. J. (2014). An integration platform for a spatially enabled society. in Steudler D. (Editor), CADASTRE 2014 and Beyond. FIG Report 61, FIG Denmark, Copenhagen, pp. 43-48.

Lüthy H. J., Kaul, C. (2015). Demands for a spatial information infrastructure fit for Cadastre 2034. FIG Working week 2015, From the Wisdom of the Ages to the Challenges of the Modern World. Sofia, Bulgaria,17-21 May 2015

Medeiros, P. (2009). Quality Assessment for Geographic Web Services (Doctoral dissertation, Master Thesis, universidade tecnica de lisboa).

NCDCDS (1988). National Committee for Digital Cartographic Data Standards, N., 1988. The proposed standard for digital cartographic data. The American Cartographer, 15: 9-140.

Wu, H., Li, Z., Zhang, H., Yang, C., & Shen, S. (2011). Monitoring and evaluating the quality of Web Map Service resources for optimizing map composition over the internet to support decision making. Computers & Geosciences, 37(4), 485-494.

W3C (2003). QoS for Web Services: Requirements and Possible Approaches. W3C Working Group Note 2003. https://www.w3.org/Architecture/qos.html, last accessed February 09 2016

BIOGRAPHICAL NOTES

Jürg H. Lüthy is member of the Management Board at SWR Geomatik AG, one of the largest geomatics companies in Switzerland. He obtained a master's degree in 1996 from Federal Institute of Technology Zurich (Switzerland) in Rural Engineering and Survey. From the same institution he holds a PhD (2007). He has many years of experience in spatial data management, transition from paper maps to data centric systems and the operation of Spatial Data Infrastructures. His current focus lies in the provision of holistic information using modern web-technologies and the transfer from practical experience with model driven SDI into the Building Information Model domain. He is the Swiss delegate to FIG Commission 3.

CONTACTS

Jürg H. Lüthy
SWR Geomatik AG
Wagistrasse 6
8952 Schlieren
SWITZERLAND
Tel. +41 43 500 44 48
Email: juerg.luethy@swr-geomatik.ch
Web site: www.swr-geomatik.ch