Spatial Data Quality: What do you mean?

Wies Vullings
Wageningen University and Research Centre
Droevendaalsesteeg 3
Wageningen, The Netherlands
Wies.vullings@wur.nl

Jandirk Bulens
Wageningen University and Research Centre
Droevendaalsesteeg 3
Wageningen, The Netherlands
Jandirk.Bulens@wur.nl

Frans I. Rip
Wageningen University and Research Centre
Droevendaalsesteeg 3
Wageningen, The Netherlands
Frans.rip@wur.nl

Martijn Boss
Wageningen University and Research Centre
Droevendaalsesteeg 3
Wageningen, The Netherlands
Martijn.Boss@wur.nl

Marcel Meijer
Wageningen University and Research Centre
Droevendaalsesteeg 3
Wageningen, The Netherlands
Marcel.Meijer@wur.nl

Gerard Hazeu
Wageningen University and Research Centre
Droevendaalsesteeg 3
Wageningen, The Netherlands
Gerard.Hazeu@wur.nl

Maarten Storm
Wageningen University and Research Centre
Droevendaalsesteeg 3
Wageningen, The Netherlands
Maarten.Storm@wur.nl

Abstract

Spatial data quality has been on the scientific agenda for a long time. Not at a priority spot, because it is a complex and not a sexy issue. However, we think that is going to change rapidly. The data explosion along with the open data policies is increasing data availability. There is finally a choice in what data to use, but how to make that choice? We see potential for spatial data quality as a selection criterion, but for it to reach its full potential more attention is needed to subjects such as determining spatial data quality, validation, communication and business case development.

In this paper we focus on the determination and communication of spatial data from a consumers as well as a producers perspective. We developed a framework in which we define different roles (consumer, producer and intermediary) and differentiate product specifications from quality specifications. We used case studies to illustrate our framework. This framework is designed following the fitness for use principle. It helps to understand the differences between producers and consumers and hence the difficulties they encounter when communicating spatial data quality. We list recommendations to improve the communication about spatial data quality between consumers, intermediaries and producers in order to improve use of spatial data and avoid capital mistakes.

Keywords: Spatial data quality, communication, producer, consumer, fitness for use

1 Introduction

Spatial data quality has been on the scientific agenda for a long time. Not at a priority spot, though, because it is a complex and not a sexy issue. However, we think that is going to change rapidly and will get a more prominent spot on the agenda. The data explosion along with the open data policies is increasing data availability. There is finally a choice in what data to use, but how to make that choice? How to be sure that we end up using the best possible data set for the application. We see potential for spatial data quality as a selection criterion. Specifically the fitness for use approach can contribute in facilitating the choice in what data to use for a specific application. In June 2014 we organized a symposium titled 'Why Spatial data quality?' More than eighty Dutch scientist and policymakers shared their thoughts on this subject. It was concluded that spatial data quality has indeed the potential to become a selection criterion and that fitness for use should be the guiding principle. However in order for it to reach its full potential more attention is needed to subjects such as determining spatial data quality, validation, communication and business case development.

It reflects that besides investing in obtaining the highest possible quality, a focus on the determination of the best fit and appropriate communication of spatial data quality will result in a better understanding of data quality and ultimately result in better value for money. Based on these findings and our broad experiences dealing with spatial data in a number of cases [1,2,3,4,5,6] we have defined a framework for spatial data quality. This framework is illustrated by case studies from a consumer as well as a producer perspective.

2 Background

According to ISO 9000 section 3.1.5 (formerly ISO 8402: 1994) quality is defined as "the totality of characteristics of an entity that bear upon its ability to satisfy stated and implied needs.” And “The purpose of describing the quality of geographic data is to facilitate the comparison and selection of the dataset best suited to application needs or requirements’ [7].

There are many standards describing quality, but since they are converging towards the ISO standards for spatial data we focus on the ISO 19157 standard describing spatial data quality by the following six groups of elements: Completeness, Logical consistency, Positional accuracy, Thematic accuracy, Temporal quality and Usability element [7].

The way these elements are defined already indicates that, besides the quality of geometric properties, also the content expressed by the thematic properties does matter. There is also a placeholder for usability although no specific properties are included in the standard since this element varies depending on the intended use.

Quality of geo-information is being assured by independent means of validation, which is a key requirement for users and producers alike for providing evidenced quality specifications [8]. Validation is a key for evaluating fitness-of-purpose of the information for a particular application, even more in the context of political reporting and decision-making [9]. Quality assured validation must follow the principles of transparency, traceability, independence, accessibility and representativeness [10].

The ISO definition stresses that the required quality is related to the intended use. For a particular use a data set can be perfect (e.g. visualisation of background in website), while it might not be acceptable for a different usage (e.g. spatial analysis). This is described as the Fitness for Use approach:- A term used to indicate that a product or service fits the customers defined purpose for that product or service [11].

Fitness for use is determined by the user perspective. For the producers perspective the slightly different approach Fit for Purpose can be used, which refers to the suitability of data for the intended use, that is, the degree to which the data meets the needs of the intended use. The difference between Fitness for use and Fitness for purpose emphasizes the fact that data providers and consumers view quality from different perspectives. Boin and Hunter [12] specify that a producer generally wants to describe how the dataset was created, whereas the consumer is likely to have questions for which he needs an answer. They find it no surprise that effective communication between them remains an issue. Also Lia et al [13] indicate that the distance between producer and consumer is increasing with acute consequences for today’s geographic information world

Boin and Hunter conclude that solutions for improved communication between producers and consumers should use the terminology of the data consumer instead of being overly technical and industry-specific. Secondly, the solution should focus on ways of describing product suitability and reliability instead of the production method [12].

The GEO-Label [14] also addresses the communication aspect by introducing a graphic means to “support efficient and effective geospatial dataset quality representation and selection on the basis of quality and fitness for use”. The emphasis of this paper is on identifying the components of fitness for use.

3 Framework

Based on experience with spatial quality projects (specifying criteria and auditing datasets) we defined a framework and we used case studies to illustrate and specify the framework. The objective of the framework is to bridge the gap between producers and consumers with regards to spatial data quality definition by improving communication at the consumer site (by specifying and elaborating the information needs) as well as on the producer site (by improving access to quality information and understanding of quality aspects of the data). (fig 1)

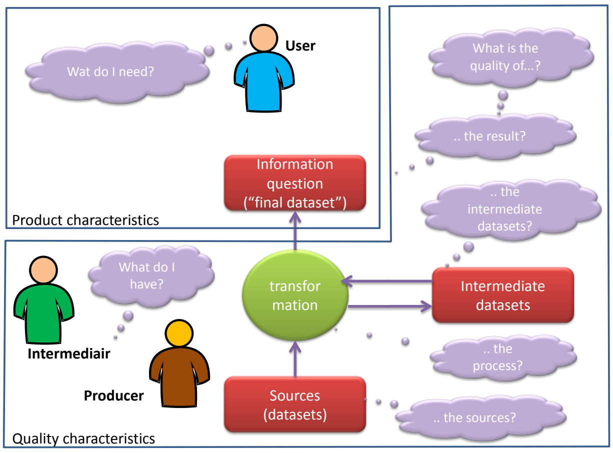


Figure 1: framework spatial data quality

In the framework the user as a consumer plays a central role, since the consumer and the context of the usage determine the necessary quality (fitness for use). By describing the use case of the consumer we identify the relevant context to be the universe of discourse. The consumer often gives spatial data quality specification within the identified Universe of discourse to his/her best knowledge, but many quality elements can be implicit and not known by the consumer. It is important to unravel the information question into criteria with the help of spatial data quality expertise. Based on this information we define the product that is wanted by the consumer. This can vary from plain data provisioning to automated procedures like an App up to providing human services. In this stage we will limit the functionality of the framework to data and the requirements of processes. We leave the quality aspect of the processes themselves as well as the institutional aspect out of our scope. They are to be included in the next stage.

When starting with the consumers information question, first all relevant product characteristics are revealed and these are related to quality requirements using the standardised quality elements. We distinguished between product characteristics and quality characteristic. For instance if the consumer is looking for data of trees in his municipality and he needs to know what type of tree is located where, a product characteristic is that the data set needs to include data on the type of tree. A quality characteristic can be that 95% of the records in the dataset provide species information.

Apart from the consumer role a producer role and an intermediary or broker role were defined. The broker is defined as a service provider of spatial data between the consumer and producer. The broker can add value to the data of the producer or supply services that provide that data as a product. Both the broker and producer should specify the product characteristics and the value of the quality characteristics of their dataset(s) and in case of data set transformations of all the ‘in between’ products as well. This information has to be comprehensible and easily accessible for the consumer. Only then a consumer can judge whether a dataset is ‘good enough’ to fit the intended use.

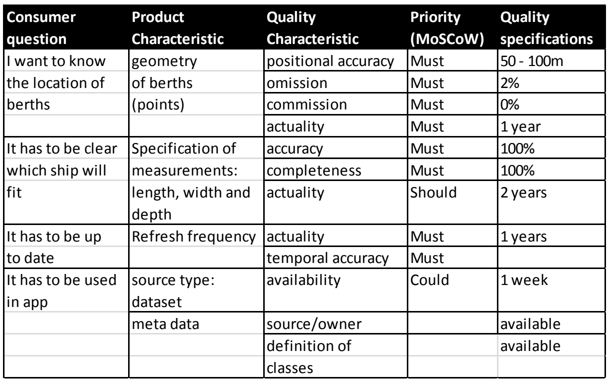
In the past producers were the ones creating datasets, usually initiated by a specific need for that data. In time more use cases can evolve that have a need for the same type of data. Common practice was to use that data or when necessary transform the data to be useful for the case. It is conceivable that a producer from a business point of view will market his data. In that case the data should be multipurpose, fitting more than one need. Instead of just one use case, the producer should make an inventory of possible use cases for which the data could be needed and for each use case the product and quality characteristics should be defined.

3.1 Case studies: Consumer perspective

Two cases are presented here to illustrate the perspective of the prospective data user. In these cases we supported consumers by unravelling their information questions into product and quality criteria. We used a simple matrix form which requires to specify per consumer question the relevant product characteristics, the quality characteristics, the priority in list and the quality specifications (quantification of the quality characteristic). Furthermore we helped them with prioritizing the criteria and quantifying them. For assigning the priority we used the MoSCoW system, in which the uppercase letters stand for Must, Should, Could and Won’t.

In the first case spatial data for the location of berths (mooring places for (river) cargo ships) in the Netherlands was asked for. When interviewing the customer it became apparent that the customer wanted to develop a planning tool for trips for shippers. The results of the case can be found in table 1. During completion of the matrix, the user himself was helped, because his question became more specific and clear, and it proved to be possible to obtain specific quality characteristics and requirements.

Table 1: Matrix elaborating consumer question location of berths



Consumer question	Product Characteristic	Quality Characteristic	Priority (MoSCoW)	Quality specifications
I want to know the location of berths	geometry of berths (points)	positional accuracy	Must	50 - 100m
		omission	Must	2%
		commission	Must	0%
		actuality	Must	1 year
It has to be clear which ship will fit	Specification of measurements: length, width and depth	accuracy	Must	100%
		completeness	Must	100%
		actuality	Should	2 years
It has to be up to date	Refresh frequency	actuality	Must	1 years
		temporal accuracy	Must	
It has to be used in app	source type: dataset	availability	Could	1 week
	meta data	source/owner		available
		definition of classes		available

The other case is about waterbodies. The consumer was interested in actual information on permanent water bodies in the Netherlands. He want it to compare the Dutch situation to other EU member states. The results of this case are in table 2.

These two cases are an indication of the multi-facetedness of fitness for use. The data properties should fit the application context of the user and it seems likely that this context varies with each user. Although the matrix helped the user to become more specific with regards to his request the whole process requires an intermediary to guide the user in making the user needs explicit.

Table 2: Matrix elaborating consumer question water bodies.

Consumer question	Product Characteristic	Quality Characteristic	Priority (MoSCoW)	Quality specifications
I need to know actual information where in the Netherlands are permanent water bodies at sufficient spatial resolution so I can compare the Dutch situation with other EU countries ?	Permanent water bodies (according to water presence indices in 2006, 2009, 2012, surface water, excluding sea and oceans	Thematic accuracy	Must	85% for area of 3 by 3 pixels
		completeness	Must	100% (no nodata)
		actuality	Must	presence on at least one reference image of every year
		logic consistency	Must	permanent waterbodies pixels can not be wetland pixels in other European Environment Agency (EEA) products
	20 by 20 m pixels (raster) Inspire grid		Must	
	covering the whole of the NL		Must	
	Projection Dutch national Grid (RDNew) and EU grid (ETRS89)		Must	
	Digital raster dataset (Geotif)		Must	
	Basic information for monitoring policy		Must	
	Inspire profile for metadata	compliant	Must	
	Describtion of followed procedure and used datasets and results	according to European Environment Agency (EEA) format	Must	

3.2 Product characteristics vs Quality characteristics

While supporting consumers by defining their desirable product characteristics and explicit their quality characteristics, it became apparent that consumers generally have similar type of product characteristics even though their information needs are very different. Furthermore these product characteristics turned out to be related to a fixed set of quality characteristics. Table 3 shows a general set of product characteristics with their quality characteristics.

This set of product characteristics with related fixed set of quality characteristics has potential to improve and facilitate the process of linking the consumers information question to quality specifications.

Table 3: general set of product characteristics with their quality characteristics

Product characteristic	Quality characteristic
Every object has to have certain properties	Thematic accuracy
	Completeness
	Temporal quality
	Lineage
	Logical consistency
Date of publication, date of data collection, update frequency	Temporal validity
	Temporal accuracy
Every object has geometry (point, line, ploygon, grid)	Positional accuracy
	Omission / commission
	Logical consistency
Type (data, service)	Availability (i.e. 24*7)
	Access (i.e. on-line)
Resolution	Positional accuracy
Meta information of the data needs specific element	Completeness
Compatible with previous versions	Lineage

3.3 Case studies: Producer perspective

The producer perspective unfolds in two ways: the producer who knows his customers and one who does not. For three Dutch datasets we investigated how the producers communicate with their consumers. We looked at their efforts to organize their consumers in user groups and to communicate with known and unknown consumers on the subject of quality of their product. We also identified potential improvements in communication.

LGN is the Land Use dataset of the Netherlands. LGN maps the entire country and is based on satellite imagery and additional data. The LGN dataset defines land use in thirty-nine classes. The dataset has been updated every 3-5 years since 1986. The newest version, LGN7, came out at the end of 2013 and documents Dutch land use in 2012 [4]. It is a commercial dataset, although negotiations on opening it up are ongoing. The dataset is mainly used by provinces, waterboards and national research institutes [14]. Its consumers are known and when a new version is available a consumer day is organized to inform the consumers of the new features and developments. They are also informed of ongoing business by mailings and via the website. Information on quality aspects are available via the website in background documentation.

A first version of the Dutch Tree register (Boomregister.nl) was created in 2013 as an experimental dataset containing canopy projection polygons for all trees in the Netherlands [5]. After that, attempts were made to validate the dataset [6] and also to find parties interested in applications or in further improvement and development. It is an open database. There is no organized user group, since it is a relatively new dataset which is generated from an experiment and without an intended customer in view. Efforts are made to find customers and gather their product and quality requirements. Quality information is available in reports [6].

	Land use (LGN)	Tree register	TOP10NL
Open/charged	charged	open	open (since 2012)
History	since 1985, new dataset every 3 years	Since 2013	since 1996 (nationwide digital data)
customer group	yes	No	yes
quality information	validation methods and results, controls and verification and metadata	Partial validations	validation methods and results, external auditreports and verification and metadata
distribution of quality information	in background document via LGN website and metadata, via user group meetings and mailings	Downloadable report (in prep.)	in background document via website and metadata, via user group meetings and mailings
Improvements in communicating with known consumers	linking potential usage to quality	Not yet	no suggestions
Increasing number of known consumers	not an issue	Not yet	registration at downloading and advertising advantages of being a known customer
Improving communication with unknown consumers	not an issue	specifically linking potential use cases with quality information via various canals (website, metadata, social media)	specifically linking potential use cases with quality information via various canals (website, metadata, social media)

Table 4:Overview of three datasets: how do the producers deal with communicating quality details concerning their dataset with known and unknown consumers.

TOP10NL is the digital topographical dataset of the Dutch Cadastre. This is the most detailed product within the national topographical base registration. It is generated from aerial photo interpretation, combined with field visits and input of other datasets [2,3]. It used to be a commercial product, but since January 1st, 2008 it is part of the Dutch system of base registrations and since January 1st, 2012 it is open for anyone to use. The original intended use was for military purposes. Five times a year a new update is available of the dataset. The aim is that any location should be updated at least every two years. The Cadastre has organised the consumers of TOP10NL in a user group that meets five times a year to discuss developments, requirements, updates etc. The user group is part of the structure of the Dutch system of base registrations and originates from the time the product was still a commercial product with known users.

In table 4 the three datasets are compared with regard to the way their producers deal with communicating quality details concerning their dataset with known and unknown consumers. Furthermore, suggestions on improving communication and increasing the number of known consumers are made per dataset. When producers know their users it is advisable to organise the user in user groups and make them part of the process where they can express their real needs. Also from this perspective the intended use is determinative for the required quality. In the process flow (fig 2) this is schematically represented.

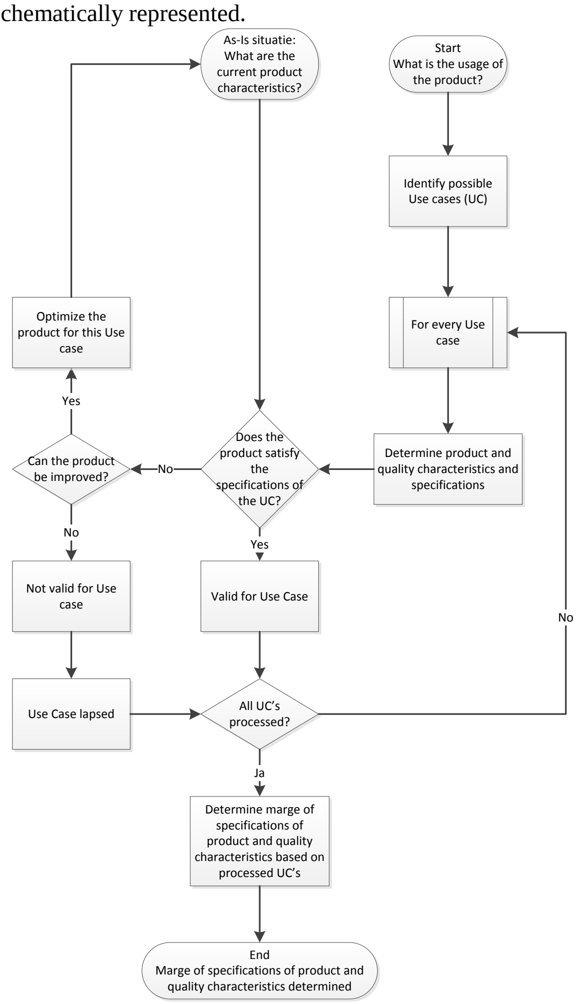

As-Is situatie: What are the current product characteristics?
Start What is the usage of the product?
Identify possible Use cases (UC)
For every Use case
Optimize the product for this Use case
Yes
Determine product and quality characteristics and specifications
Does the product satisfy the specifications of the UC?
No
Can the product be improved?
No
Yes
Not valid for Use case
Valid for Use Case
No
Use Case lapsed
All UC's processed?
Ja
Determine marge of specifications of product and quality characteristics based on processed UC's
End Marge of specifications of product and quality characteristics determined

Figure 2: process flow of identifying customer product and quality requirements by producer

In case of open data, not all users are known beforehand. Producers can try to get to know users by suggesting voluntary registration at downloading and try to involve those users in their user groups. However there will always be a group of unknown users. For those unknown consumers, producers should try to make the quality information available. It should preferably be linked to possible applications so consumers can relate to their own intended use and get information on the fitness of the dataset for their intended use. The information should be as easily accessible as possible by communicating it via websites and social media and metadata.

4 Discussion and future research/ recommendations

In order to use spatial data quality as a selection criterion when choosing a data set for usage in an application, the determination and communication of spatial data quality between consumers, brokers and producers needs to improve. Suggestions:

For consumers: Facilitate a mediator to support consumers in specifying product and quality criteria for their information needs in order to find best match. Dialogue is needed and it is not easy to facilitate this without dialogue, e.g. with a wizard or check list. The list of product specifications with related quality specification can be the used in the dialogue as a guideline to translate the information question into a set of product and quality specifications.

For producers: communicate with consumers if possible, organize them in user groups and gather use cases to improve your products. If consumers are not known, try to find out who they are (registration at downloading), so they can be contacted. In case of an open product there will always be unknown users. Producers should anticipate and communicate with them about the spatial quality of a product in a comprehensible and accessible way (metadata, internet, social media). Emphasis should lie on the combination of possible use cases and spatial quality.

In the framework presented here we focussed firstly on extracting and defining characteristics based on its intended use. Future work will be to expand the framework to other relevant quality properties of data. One extension will be to assess the information published besides the data itself. One can think of the availability of feature catalogues containing commonly, standardized and excepted definitions of spatial features and their attributes. Another could indicate the level of compliance with existing standards (INSPIRE). Other options are proper documentation and metadata using standards, availability of managed code lists accessible through registries based on described standardized hierarchies as for example SKOS. Also sharing licenses like Creative Commons (is this data open or with certain restrictions and costs?). Furthermore, we like to continue focussing on the communication aspects, so all parties involved can find and know what is meant by quality information.

It all matters when one has the luxury to choose what data to use. This framework will increase the use of spatial data and help to avoid capital mistakes.

References

[1] Meijer, M. and L.A.E. Vullings, 2012. Kwaliteit van ruimtelijke data in relatie tot het LPIS; kwaliteitsaspecten rondom het beheer van ruimtelijke data. Wageningen, Alterra, Alterra-rapport 2285.

[2] Storm, M.H., M. Knotters and D. Brus, 2012. Controlemethodiek Basisregistratie Topografie. Wageningen, Alterra.

[3] Storm, M.H., M. Knotters, D.J. Brus, 2012. Audit Basisregistratie Topografie, resultaten van een eerste wettelijk vereiste externe controle op de kwaliteit van de BRT.

[4] Hazeu, G.W., C. Schuiling, G.J. Dorland, G.J. Roerink, H.S.D. Naeff and R.A. Smidt, 2014. Landelijk Grondgebruiksbestand Nederland versie 7 (LGN7); Vervaardiging, nauwkeurigheid en gebruik. Wageningen, Alterra Wageningen UR (University & Research Centre), Alterra-rapport 2548. 86 blz.; 16 fig.; 12 tab.; 15 ref.

[5] Rip, F.I. and J. Bulens. IM-Tree. Towards an information model for an integrated tree register. 16th AGILE conference on Geographic Information Science, 14-17 May 2013, Leuven Belgium.

[6] Meijer, M., F.I. Rip, R. van Benthem, J. Clement, and C. van der Sande. Boomkronen afleiden uit het Actueel Hoogtebestand Nederland. Alterra rapport (in prep.). Wageningen University and Research Centre, 2015.

[7] ISO 19157:2013 Geographic information - Data quality

[8] Justice, C.O., Belward, A., Morisette, J., Lewis, P., Privette, J. & Baret, F., 2000, Developments in the 'validation' of satellite sensor products for the study of land surface. International Journal of Remote Sensing, 21, 3383-3390

[9] Congalton, R., 2001. Accuracy assessment and validation of remotely sensed and other spatial information. International Journal of Wildland Fire, 10, 321-328

[10] The Quality Assurance Framework for Earth observation (QA4EO - http://qa4eo.org/background.html)

[11] American Society for Quality, 23 September 2010 12:40:01, http://www.asq.org/glossary/ External

[12] Boin, A.T. and G. J. Hunter. 2006. Do spatial data consumers really understand data quality information? 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Eds. M. Caetano and M. Painho.

[13] Deren Lia, Jingxiong Zhangb, Huayi Wua, 2012, Spatial data quality and beyond. In International Journal of Geographical Information Science Vol. 26, No. 12, December 2012, 2277–2290

[14] Lush, V., L. Bastin, J. Lumsden, 2012: Developing a GEO Label: Providing the GIS Community with Quality Metadata Visualisation Tools

[15] Oort, P.A.J. van, G.W. Hazeu, H. Kramer, A.K. Bregt, F.I. Rip, 2010: Social networks in spatial data infrastructures. GeoJournal (2010) 75:105-118