OCENA PRIMERNOSTI UPORABE PROSTORSKIH PODATKOVNIH NIZOV

GEOGRAPHICAL DATA SETS FITNESS OF USE EVALUATION

Elżbieta Bielecka

UDK: 556.043:659.113.21
Klasifikacija prispevka po COBISS.SI: 1.04
Prispelo: 24.12.2014
Sprejeto: 11.5.2015

DOI: 10.15292/geodetski-vestnik.2015.02.335-348
PROFESSIONAL ARTICLE
Received: 24.12.2014
Accepted: 11.5.2015

SI | EN

IZVLEČEK

Številčnost dostopnih prostorskih podatkovnih nizov je pogosto povezana s težavo, kako naj uporabnik izbere najprimernejše podatke. Teoretično bi morali biti v pomoč pri izbiri primernih podatkov metapodatki, toda številne študije kažejo, da je koristnost metapodatkov različnih ponudnikov za povprečnega uporabnika omejena. Namen prispevka je predstaviti ključne informacije, ki jih uporabnik potrebuje pri izboru najprimernejših prostorskih podatkov za neki namen. Te informacije smo primerjali z informacijami, ki jih lahko uporabnik pridobi z metapodatki, prek opisa podatkov in s samimi podatki. Pomemben rezultat raziskave je nabor kakovostnih parametrov in tudi informacij, ki jih uporabniki pogrešajo pri metapodatkih. Uporabniki so izpostavili, da se zara-di nepopolnosti podatkovnih nizov precej zmanjšata informativna vrednost in primernost uporabe podatkov. To smo ugotovili tudi za podatke o stavbah, ki so del poljske državne topografske zbirke. Rezultati raziskave kažejo, da je kakovost podatkov raznolika v sami podatkovni zbirki in zahtevana kakovost nekaterih podatkovnih podnizov ni zagotovljena. V sklepu tako podajamo predlog za dva dodatna kakovostna podelementa – neobvezen opisni podatek in manjkajočo vrednost – ki bi lahko prispevala k lažji oceni primernosti uporabe podatkov za posamezen namen.

KLJUČNE BESEDE

kakovost podatkov, prostorski podatki, metapodatki, primernost uporabe, topografska podatkovna zbirka

ABSTRACT

The large number of commonly available geographical data sets means that users of this data face a difficult choice in selecting the set that best meets their requirements. In theory, metadata is helpful in this, but many studies suggest that the metadata created by data producers is incomprehensible to average users. The article aims to identify the essential information that users need to acquire the geographical data set that fits their needs. This information is then compared with the information that users can obtain from metadata, product data specifications, and the data itself. As a result of a survey (the most important data quality elements were identified, as well as some information pertinent to users that is missing in metadata. The users stressed that a lack of value for optional attributes considerably decreases the informative value and fitness for use of existing data sets. This was also observed while analysing the building thematic data layer, which is a part of The Polish National Topographic Database. The research shows that data quality is diversified within a database, and it may happen that for some subsets of data, quality criteria are not met. Finally, two data quality subelements – optional attribute and void value - were proposed, which will overcome some difficulties in assessing the fitness for use of data.

KEY WORDS

data quality, spatial data, metadata, fitness for use, topographic database

1 INTRODUCTION

For several years we have observed a dramatic increase in the geographical data available on websites. In many European countries, this increase is associated with the establishment of the infrastructure for spatial information, INSPIRE. INSPIRE obliges Member States of the European Union to reuse collected data, for instance, via network services enabling the search, viewing and downloading of geographical data (Directive 2007/2/EC). Universal data access means that more and more users of spatial data face the difficult choice of selecting a set that most closely meets their requirements. This problem, defined as 'fitness for use', has been widely discussed in literature for more than 40 years. Fitness for use, described as the extent to which a product best serves the purposes of the user, was introduced by Juran (1974), and popularised in GIS by Chrisman (1983). It has been widely used ever since, because it takes into account customer intentions for use of the product, instead of focusing on conformance to technical specifications. According to Juran (2010), fitness for use is a certain kind of connection between data quality and the user. Although fitness for use captures the essence of data quality, it is difficult to measure quality using this broad definition (Kahn et al. 2002). The determination of fitness for use of a data set relies on knowledge and an individual's expertise. Each user group has particular requirements, so different aspects of usability have to be considered. The fitness for use decision can be easily determined if the user's quality requirement is known. Therefore, gathering information about users and grouping them according to their behaviour is of the utmost importance. This was the aim of a study conducted by Boin and Hunter (2008), who grouped users according to their background (e.g. architect, social researcher, acoustic analyst, cartographer, archaeologist, technician, etc.).

Fitness for use is the ability of the data set to fit the stated user requirements and application specifications. Based on that, Redman (2000) suggested that a data set that is fit for use should be accessible, accurate, timely, complete, consistent with other sources, relevant, comprehensive, provide a proper level of detail, be easy to read and easy to interpret. Frank et al. (2004) described a procedure of selecting the best data set for a given task based on a quantitative assessment of spatial data sets. The disadvantage of this procedure was that it was elaborated only for one user group – pedestrian tourists. Wright (2006) described a model for fitness for use for the support of military decisions which included both error distribution and spatial relationships. Paradis and Beard (1994) drew up a data quality filter to efficiently communicate data quality to a decision maker. Devillers et al. (2006, 2010) developed and implemented a system for data quality assessment called the Multidimensional User Manual (MUM). The system allows the management of geographical data quality as well as the communication of information on the quality of indicators used, at different levels of detail. Bielecka et al. (2014a) suggested that information on data quality should be presented in the form of a chropleth map, and argued that this way of reporting quality is better understood than metadata.

In practice, as was stated by Wright (2006), data producers, data providers or data custodians issue a disclaimer to the effect that "determining fitness for use is solely the users' responsibility". Producers of geographical data assume that users are able to determine the fitness for use of a geographical data set before they use the data set. They expect users to look at the description of a data set contained in a metadata file, and compare it with the list of quality requirements. Thus, the user is expected to understand the characteristics of a given data set and the extent of its potential use solely based on metadata. Metadata defined as "data on data" (ISO 19 115:2014) should be elaborated in a standardised way. A conceptual model for describing digital geographic data provides the ISO 19 115: Geographic information - Metadata – Part 1: Fundamentals. It defines metadata elements, and establishes a common set of metadata terminology, definitions, and extension procedures. When implemented by a data producer the standard facilitate data discovery, retrieval and reuse. It enables users to determine whether geographic data, available to the public use, will be of use to them. The ISO 19115 standard defines an extensive set of metadata elements, however only a subset of elements is used. The minimum set of metadata required to serve the full range of metadata applications (e.g. data discovery, determining data fitness for use, data access, data transfer, and use of geographic data) is known as "core metadata". The core metadata includes 22 metadata elements, out of which 7 are mandatory, 11 – optional and 4 – conditional. The core metadata provides information on several aspects of the data sets, such as the identification and classification of geographical data, keywords, spatial reference system, geographical location, temporal reference, data quality and validity, restrictions related to access and use, as well as organisations responsible for the establishment, management, maintenance and distribution of geographical data sets. Using the core metadata elements increases interoperability, allowing users to understand without ambiguity the geographic data and the related metadata provided by either the producer or the distributor. All dataset metadata profiles of the ISO 19 115 shall include these core metadata elements.

Metadata provides resource characteristics that can be queried and presented for evaluation and further processing by humans and computers. Metadata, however, are prepared by data producers who describe the set with regard to its conformance with data specifications. Conformance to specifications measures how well the data meets the targets and tolerances determined by its designers. Therefore, knowledge of the data specification is essential to understanding the thematic scope, level of detail, and the quality of the data. However, although conformance to specifications is directly measurable, it is rarely directly related to the consumer's understanding of quality.

A common method to describe, manage, and present the quality of geographical data sets is provided in the ISO 19157:2013 Geographic information - Data quality, which replaced the international standards on geographic information ISO 19113:2002 – Quality principles, ISO 19114: 2003 – Quality evaluation procedures, and ISO/TS 19138:2006 – Data quality measures. The objective of the ISO 19 157 is to guide the data producer in choosing the appropriate data quality measure, and the user in the evaluation of the fitness for use of a data set. This is achieved by standardising the components and structure of data quality elements as well as defining commonly used measures for assessing data quality. It should also be mentioned that none of the ISO data quality standards attempts to define minimum acceptable levels of quality for geographical data.

Information on data quality is reported in a metadata file, elaborated according to ISO 19115. The profile of ISO 19115 for the European infrastructure of geographical information, INSPIRE, is provided in COMMISSION REGULATION (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council with respect to metadata (CR, 2008). In this document "quality" means the totality of characteristics of a product that bear on its ability to satisfy stated and implied needs. The detailed description of the contents related to the data quality in the INSPIRE Directive, regulations for its implementation and guidelines, as well as the requirements related to the quality of geographical data is given by Ažman (2011).

Poland as a Member State of The European Union also implements the INSPIRE metadata profile for describing information about national geographical data sets; it is called *POLISH METADATA PROFILE FOR GEODETIC AND CARTOGRAPHIC RESOURCES* (Baranowski et al. 2012; Bielecka 2007; MAiC, 2013). This profile includes all core metadata elements as well as information on resource type, unique resource identifier, keywords, conformity (including document names and level of conformity), condition for access and use, and limitation on public access, all together 29 metadata elements (CR, 2008).

Nevertheless, all benefits of using metadata, studies conducted by Delavar et al (2010), Devillers and Bédard (2010), and Goodchild (2008) show that the majority of users do not understand the content of the metadata. This leads to the paradoxical situation that on the one hand a user has easier access to more geographical data than ever before, but on the other hand he knows less about this data and its quality.

The research aims to identify the essential information that users need to acquire the geographical data set that fits their needs. This information is then faced with the information that users can obtain from metadata, product data specifications, and the data itself, based on the data quality and thematic scope analysis of a chosen geographic data set. Finally, two new data quality elements were proposed to facilitate data suitability evaluation. This is the first step to an in-depth study of elaborating quality indicators for geographical data to enable comparison of datasets against user requirements as well as to convey quality information about geographical data in a clear manner.

2 METHODS AND DATA

2.1 Main research assumptions

The research was conducted in two main stages. In the first stage, users identified the information they required to choose the best data for performing a given task. It was done on the basis of a survey conducted among 594 people, including 350 representatives of local self-government administration, 144 officials of the central administration, 60 final year students of geography or geodesy and cartography, and 40 GIS researches. All respondents just completed 120 hours training on INSPIRE and in the use of geographic information, metadata and discovery services of geographical data sets, as well as geographical data quality assessment in the context of their suitability for a given task. These geographical data users were asked to describe how they discovered the best data to perform a given task and to identify the metadata elements of the Polish Metadata Profile for Geodetic and Cartographic Resources that are essential for data discovery and fitness for use evaluation. The survey participants were furthermore required to specify the supplementary information that should be reported in metadata to facilitate data discovery and suitability assessment; and how the metadata report could be improved to become more user friendly, i.e. easy to understand.

In the second stage of the study, the quality of the subset of geographic data set was evaluated. The results of this evaluation were then compared with the information about quality stored in metadata and the quality criteria published in data specifications. The purpose of this phase was to examine whether the quality of any part of the data set (or database) is in line with the quality requirements set up for the entire data set as well as to test whether the complaints of geographic data users on inconsistent quality of data sets are justified. The analysis was conducted for the *Building* thematic data layer, which is a part of the Polish topographic database. The following reasons justified this decision:

- buildings and built-up areas are of interest to many research domains e.g. cartography (Steiniger et al. 2008), photogrammetry (Zhang et al.2013; Zhang et al. 2010; Rau, Chen 2003), geodesy (Bielecka et al. 2014b), urban geography as well as spatial and physical planning (Montero et al. 2010), disaster damage assessment (Takashima et al. 2003);
- information on building locations and characteristics is of utmost importance to economic sustainable development (Li et al., 2014) and is essential for public administration and citizens (Frank et al. 2004);
- they are reference data for other data (e.g. address points) (D2.8.III.2, 2013).

The analysis was carried out at the level of data set, objects (features), and attributes; and included:

- The completeness of the buildings, which was established by comparing the number of buildings, gathered in the *Building* set and the Polish cadastral data (here adopted as a reference data).
- The completeness of the attributes and values of 'null reason', analysed by the data filtration method.
- Conceptual consistency expressed by the following rules:
    - The accuracy of the allocation of the address point, analysed by the spatial query method.
    - The fulfilment of the condition of the minimum surface and the minimum length of sides, referred to as a result of data filtration.
- The accuracy of defining the location referred to by measuring the displacement of a building in relation to its location in the cadastral database, and expressed by the mean absolute error (MAE) and standard deviation (RMSE). This positional accuracy analysis covered 5% of buildings stored in the database, spread evenly over the entire area of the study.

The buildings attributes of completeness and conceptual consistency were evaluated using an internal direct evaluation method. The positional accuracy and the buildings completeness were estimated by an external direct evaluation method.

2.2 Data and study area

2.2.1 Building thematic data layer

The subject of quality assessment is the *Building* thematic data layer, which is one of the data sets of the Database of Topographic Objects (BDOT10k), maintained by the Head Office of Geodesy and Cartography in Poland. BDOT10k is a spatially continuous, vector database with the thematic scope and a level of detail corresponding to contemporary, civilian topographic maps at a scale of 1:10,000. In accordance with the specifications of the data (Surveyor General of Poland, 2003), buildings are represented geometrically by polygons (defined by a ground level outline) and described by 25 attributes, of which 11 are of a technical nature, and contain, among other things, information about the data creator, system ID etc. The remaining 14 attributes describe the characteristics of the building, where 10 are mandatory (M), and 4 optional (O). A list of the thematic attributes is given in table 1.

Table 1: Attributes assigned to buildings (Source: Surveyor General of Poland, 2003).

	Attribute name*	Status**	Description
1	Id	M	Unique object identifier
2	CurrentUse	M	activity designated for the building classification of buildings, based on their current use
3	DetailedCurrentUse	M	
4	BuildingName	C	name of the building (if exist)
5	NoOfFloors	O	Number of floors
6	HeightAboveGround	O	In meters
7	monument	M	Boolean value, 1 – if a building is a monument; 0-if a building is not a monument
8	temporalValidity	M	Date of modification
9	PositionalAccuracy	O	The estimated absolute positional accuracy of the (X,Y) coordinates of the building geometry
10	PositionalAccuracyCategory	M	1 – precise; 2 – approximate; 3 – imprecise
11	GeometryDataSource	M	Field measurements, base map, ortophotomap, thematic map, other
12	AttributeDataSource	M	Field measurements, base map, ortophotomap, thematic map, other
13	Geometry	M	GM_Surface
14	ObjectState	M	Utilized, under construction, destroyed, provisional

*The names of attributes were translated to English, originally they were in Polish

**M-mandatory, O-optional, C-conditional

According to data specification, the database should contain 98% of buildings, created on the basis of the vectorised orthophotomaps, the Polish base map at the scale 1:500-1:1,000 or the cadastral data. Each building, with the exception of outbuildings, should be assigned an address.

2.2.2 Cadastral data

The Register of Lands and Buildings, Polish cadastre, is a uniform collection of systematically updated data on lands, buildings and premises, their owners and other parties holding these lands, buildings and premises as defined in Geodetic and Cartographic Law (2014). The Register covers the whole territory of Poland (312,000 km^2, and 33 millions of parcels). Data are open to public and commonly accessible at the district level. Buildings are defined by a ground level outline and geometrically represented by a polygon. Each building is characterised by several attributes (e.g. address, date of construction, number of floors, type of building, technical standard, etc.). The location of a building is determined through field measurements, using surveying methods and techniques, with a positional accuracy of 0.10 m. Hence, in this study, the cadastral data is adopted as reference data.

2.2.3 Study area

The study site for this research is the city of Mazovian Minsk (depicted in Fig.1). It is a town in central Poland, with 39,299 inhabitants covering an area of 11.61 km^2, located 38 km east of Warsaw, the capital city of Poland. The *Building* thematic data layer of Mazovian Minsk includes 7476 buildings. Some characteristics of data used for the study are summarised in Table 2. Data are described by metadata (only in Polish), which are available from geoportal.gov.pl metadata Catalogue Services for Web.

Table 2: Descriptive characteristics of used data sets.

Item	*Building* thematic layer	Cadastral data
Geographical extent:		
westbound longitudes	E 21° 30′ 30′′	
eastbound longitudes	E 21° 36′ 22′′	
southbound latitudes	N 52° 09′ 45′′	
northbound latitudes	N 52° 11′ 52′′	
Updateness	2012	
Area covered [km^2]	11.61	
Corresponding map scale	1:10,000	1:1000
Source of geometry	Ortophotomaps or analogue topographic map vectorisation	Field measurements or ortophotomaps vectorisation
RMS of building location [m]	5	0.10
Number of buildings	7476	7561

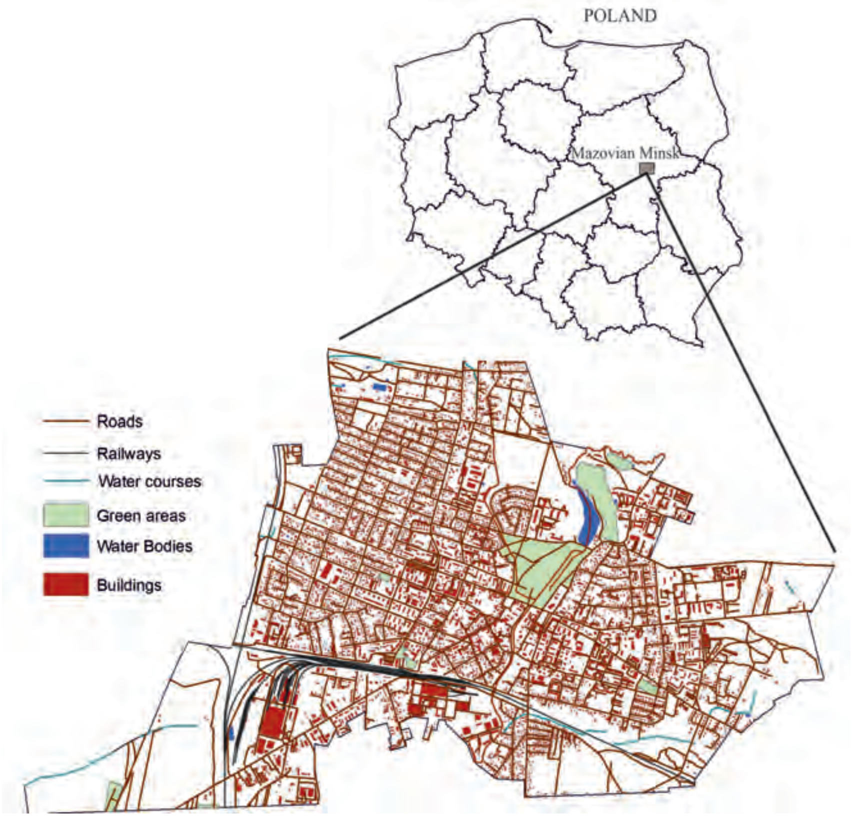

Figure 1: Geographic location of the area of interest.

3 RESULTS AND DISCUSSION

3.1 Results of the survey

Based on the survey results it was possible to identify the metadata elements most often used for data discovery and evaluation. Moreover, the user desired data quality components to be included in the standard metadata report were listed.

The most frequently searched geographical data informational attributes, stored in metadata, include: geographical location and extent, thematic scope, spatial resolution, distribution format, access and use restrictions, the organisation responsible and its reliability. Survey respondents also considered the lineage metadata element as useful much of the time. The geographical data sets users underlined their interest in good quality metadata. The people surveyed stated that complete, well documented metadata records are essential while assessing geographic data quality. They complained about often incomplete metadata records. In particular, completeness (including completeness of objects attributes), positional accuracy, restrictions related to access and use, and lineage are typically missing as regards the geographical data they have encountered. A suggestion about a better, more standardised way of lineage metadata formulation, including detailed information about data processing steps, was made. About 50% of respondents pointed out that the licensing information is nearly always missing. This leads to the conclusion that even though the ISO provides clear methodology for metadata elaborating and publishing, they are inconsistently used. This was also demonstrated by Boin and Hunter (2006) and Lush et al. (2012).

More than 50% of those polled stated that they generally rely on peer recommendations when selecting a dataset. 40% of respondents rely on their personal familiarity and reliability of the data provider. While the state data provider is perceived as the most reliable. A small number of respondents (less than 10%) use metadata for searching geographical data. From the surveyed geodata users' perspective some data quality components are missing (in the standard metadata report) to clearly define the quality and suitability of the given data set. They explicitly listed: presence of "null reason" attribute values, names of optional attributes (and therefore often not having their values), and information on positional accuracy. The poll participants mentioned that the way metadata reports are formulated is difficult to analyse and understand. Effective visualisation methods for metadata records were studied in depth (Devillers and Bédard, 2010; Devillers et al. 2006; 2010; Ivánová et al., 2013) but proposed solutions are still in the prototype phase. The respondents emphasised other important aspects of potential metadata visualisation, namely the possibility of evaluating the spatial distribution of incomplete data (e.g. places where missing or excess objects are located), and the ability to compare the quality of different data sets.

Based on the survey and some literature (Frank et al. 2004; Boin and Hunter 2008; Bielecka 2010; Lush el al. 2012; Ivánová et. al. 2013) it is possible to recapitulate the information needed by the user to discover and evaluate the geographic data set and compare this information with the Polish metadata profile (table 3).

Table 3: Comparison of the surveyed users' needs and the Polish metadata profile.

Information needed by a user	**POLISH METADATA PROFILE element name**	**Comments**
Geographical extent of the data set	Geographic Bonding Box	Expressed in longitude and latitude in decimal degrees
Thematic scope of the data set	Resource Abstract Keyword Topic Category	This information is directly available in the resource abstract, and indirectly via keywords and topic category
Spatial resolution	Spatial Resolution	Expressed in the scale denominator (of a comparable hardcopy map or chart) or ground sample distance.
Update	Temporal Reference	It is mandatory to fill one of the four subelements: temporal extent, date of publication, date of last revision, date of creation. The information about update is provided by the last revision date, however this sub-item may not be supplied.
Positional accuracy	-	Not covered by the Polish profile of metadata, however some information on positional accuracy is available indirectly by Spatial Resolution metadata element.
Lineage	Lineage	General lineage statement is mandatory.
Completeness	(indirectly by Conformity)	It is assumed that information on data completeness could be available through metadata element Conformity, which shows the degree of conformity with the implementing rules on interoperability of spatial data sets. One of the rules addresses the completeness of data set.
Completeness of features	-	This information is not available in Polish and INSPIRE metadata.
Completeness of attributes	-	This information is not available in Polish and INSPIRE metadata.
Presence of 'null reason' attributes value	-	Information not covered by ISO 19 115.
Variable accuracy within the set	–	Information not covered by ISO 19 115.
Restrictions related to access and use: conditions applying for access and use; limitations on public access	Condition for access and use Limitation on public access	The information is mandatory.
Organisation responsible	Responsible Organisation	Mandatory information contains the name of organisations and a contact e-mail address.

3.2 Results of Mazovian Minsk Building data layer quality analysis

The Mazovian Minsk building completeness analysis revealed the differences between the buildings data stored in the BDOT10k and the reference data set, i.e. Polish cadastral data, namely1.9% data omission and 0.8% data addition. These errors, however, are within the data specifications, and the missing buildings are randomly dispersed in the built-up area.

Some attributes with missing values were found during their completeness control. However, all attributes having null values were optional, namely: *Positional Accuracy*, *Number of Floors, Height Above Ground*, and *Building Name* (see tab.1).

Conceptual consistency is defined as adherence to the rules of the conceptual schema related to the fulfilment of the condition of minimum surface and minimum side length and the accuracy of the address point assignments. More than 4.7% of buildings fail to meet the criterion of minimum building area size, and 10% of features fail to meet the minimum side length criterion, which exceed the BDOT10k data specifications. These results reflect the issues with the data generalisation one of the most difficult tasks in data collecting that in the case of Mazovian Minsk buildings was performed incorrectly for 10% of the analysed objects. While address points were incorrectly assigned only in an insignificant 0.33% of all cases.

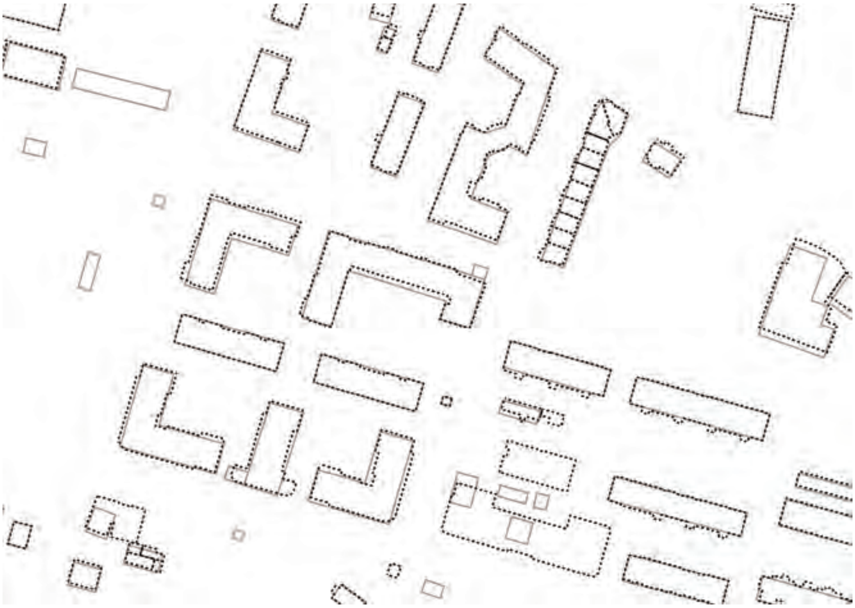

Figure 2. Displacement of geographical location of buildings stored in BDOT10k (solid, grey line) in comparison to cadastral data (dashed, black line)

The mean absolute error (MAE) of the location of the examined buildings with respect to their location in the reference data set, calculated on the basis of 5% of the sample data, amounts to 1.4 m (with a maximum of 10.1 m). The positional accuracy of buildings in the BDOT10k is expressed in a descriptive way, i.e. *Positional category,* which can take one of the following values: 'precise', 'approximate', or 'imprecise'. The accuracy of 98% of the analysed buildings was described as 'precise'; the rest (135 items) was given 'approximate' attribute value. The location MAE of buildings for which the category of accuracy is specified as 'precise' is 1.3 m, while for buildings with an approximate accuracy category is 2.0 m, both well within the data specifications. Determining the accuracy of the location of the buildings in a descriptive way is incompatible with the quality assessment methodology defined in ISO standards, but in general it is easier to understand by users. A sample offset position of the buildings is shown in Figure 2.

No spatial dependence was observed between objects displaying a less accurate location and a lack of attribute values. Spatial distribution of those objects with a probability of 0.05 is random. Moreover, the correlation analysis showed that the building positional accuracy does not depend on methods of geometry capturing , it rather depends on the object state, namely: in use, under construction, destroyed or temporary.

3.3 Discussion on Mazovian Minsk data quality

Table 4. Results of evaluation of data quality and information on data quality found in metadata and data specifications.

Data quality element	*Metadata*	*Data specification*	*Results of Mazovian Minsk Buildings layer analysis*
Completeness of features:			
— Omission	no information	2%	1.9%
— Commission	no information	no information	0.9%
Completeness of attributes	no information	Values of all mandatory attributes have to be given	Missing only values of 3 optional attributes
Conformity			
— Degree of conformity	conformant	conformant	Non conformant
Conceptual consistency:			
— Min. building area (40 m^2)	no information	1%	4.7%
— Min. polygon side (4 m)	no information	0%	10%
— Spatial relations with address point	no information	0%	0.33%
Positional accuracy	no information	Only descriptive information – precise, approximate, imprecise	Descriptive: Precise, approximate, imprecise. Mean displacement 1.4 m RMSE 0.73 m
Spatial resolution	1:10 000	1:10 000	1:10 000
Lineage	General information about data sources of and its spatial resolution	Not applicable	Not tested

The analysed *Building* thematic data layer of Mazovian Minsk is a part of the Database of Topographic Objects - an official Polish database. The quality of the BDOT10k was evaluated by the Marshal's Office of Geodesy and Cartography (data producer and provider). And the data is conformant with data specifications established by The Surveyor General of Poland (2003).

The chosen data quality analysis showed some discrepancies between the data quality values stated in the BDOT10k data specifications and the quality values obtained for the *Building* thematic data layer of Mazovian Minsk area (table 4). It illustrates the variability of the data quality within one database. This example also confirms the previously mentioned remark that data specifications together with metadata are often insufficient for data suitability evaluation.

3.4 Proposition of new data quality elements

The deeper analysis of the user requirements, the data quality evaluation results, data specifications and metadata (see tab. 3 and 4) shows that not all the desired information for the data suitability assessment can be found in their metadata, and some of it is also absent in the data specifications. In particular, information about attributes completeness (ex. 'null reason' values presence) and information about the variable data quality within the set is still missing.

Based on the results of the survey and the Mazovian Minsk *Buildings* data analysis, the two new data quality subelements are proposed, both belonging to the data and attributes completeness element:

- *Optional attribute* – indicating the presence of optional attributes (therefore often not having their values); expressed as an overall number of optional attributes or a ratio of optional attributes to all attributes.
- *Void values* – indicating the presence of voidable attributes; expressed as an overall number of "null reason values" or a ratio of "null reason values" to all attributes values.

Furthermore, the information about the spatial distribution of incomplete data (i.e. missing objects or voidable attributes) could be described in the lineage data quality element. Such distribution can be expressed using one of the spatial statistics indexes, e.g. Moran I spatial autocorrelation or Getis –Ord General G statistics.

An efficient way to communicate information about spatial data quality is a thematic map. An example of such a solution with regards to the location accuracy is presented in the work of Devillers et al. (2006, 2010), and with regards to the attributes completeness is proposed in Bielecka et al. (2014a).

4 CONCLUSIONS

Answering the main research question as to whether geographical data sets metadata is good enough to evaluate the data fitness for use, the surveyed data users concluded that it is not. Firstly, there are missing some data quality components in the metadata standard report, i.e. information about the presence of voidable attribute values, names of optional attributes, and detailed information on positional accuracy. Secondly, they recall some difficulties with the understanding and exploitation of the metadata standard reports for data suitability analysis. Thirdly, the pooled geographical data users complained about encountering incomplete metadata records.

According to ISO "any description of reality is always an abstraction, always partial, and always just one of many possible views" (ISO 19103:2005). This was also confirmed by the results of the second part of the described research, namely the data quality and thematic scope analysis of a chosen geographic data set piece with emphasis on features and their attributes completeness, conceptual consistency and positional accuracy. The analysed data set is characterised by the relatively good and uniform geometric quality, and the satisfactory data and their attributes completeness (except for three optional attributes). The Mazovian Minsk buildings data conceptual consistency was less sound but still tolerable. This analysis confirmed the surveyed geodata users outlined the lack of some users desired data quality elements. Answering this issue, the author of this paper proposes the extension of the set of data quality elements by two subelements for data attributes completeness: one indicates the presence of optional attributes and the other deals with the presence of voidable attributes.

The study shows that the quality of geographical data is not uniform and the metadata that describes the data is insufficient to assess the quality of the data by the user. Users are expected to understand the characteristics of a given data set and the extent of its potential use from the metadata. However, there is still a gap between what the quality assessment mapping experts can produce and the information that users can understand and use. The suitability of a dataset does not automatically follow from the data quality description, reported in the metadata. Metadata provides information to enable a user to ascertain that data fit for a given purpose exists, and to evaluate its properties in a general way. After the discovery of a data set, more detailed information about individual data sets is needed, and more comprehensive and more specific metadata is required. The proposed extensions of the set of data quality elements will overcome some difficulties in assessing the fitness for use of data.

ACKNOWLEDGEMENTS

This paper was prepared as an outcome of statutory research no. PBS/854/2013, conducted at the Military University of Technology, Faculty of Civil Engineering and Geodesy, Institute of Geodesy.

References

Ažman, I. (2011). Data quality and the INSPIRE Directive. Gedetski vestnik, 55 (2), 194–204. DOI: http://dx.doi.org/10.15292/geodetski-vestnik.2011.02.193-204

Baranowski, M, Kmiecik, A., Gasiorowski, J., Bednarek, M. (2012). An INSPIRE Metadata Regulation Implementation – Learning Experience. Paper presented at the INSPIRE Conference, 23–27 June 2012, Istanbul, Turkey.

Bielecka, E. (2007). INSPIRE metadata implementing Rules. Archives of Photogrammetry, Cartography and Remote Sensing, 17a, 43–52.

Bielecka, E., Leszczynska, M., Hall, P. (2014a). User perspective on geospatial data quality. Case study of the Polish Topographic Database. In The 9th International Conference "ENVIRONMENTAL ENGINEERING" 22–23 May 2014, Vilnius, Lithuania, selected paper. DOI: http://dx.doi.org/10.3846/enviro.2014.193

Bielecka, E., Pokonieczny, K., Kamiński P. (2014b). Study on spatial distribution of horizontal geodetic control points in rural areas. Acta Geodaetica et Geophysica, 49 (3), 357–368. DOI: http://dx.doi.org/10.1007/s40328-014-0056-6

Boin, A. T., Hunter, G. (2008). What Communicates Quality to the Spatial Data Consumer? Quality Aspects in Spatial Data Mining. CRC Press, pp. 285–296.

Chrisman, N. R. (1983). The role of quality information in long-term functioning of a GIS. Proceedings of AUTOCART06, 2, 303–321. Falls Church: ASPRS.

CR. (2008). 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. L 326/12.

D2.8.III.2 (2013). D2.8.III.2 INSPIRE Data Specification on Buildings–Technical Guidelines.

Delavar, M., Devillers, R. (2010). Spatial data quality: From process to decisions. Transactions in GIS, 14 (4): 379–386.

Devillers, R., Bédard, Y. (2010). Responsible use of geospatial data - Have you thought recently of your responsibility for the consequences of your actions? GEOconnexions International Magazine, June 2010: 30–32.

Devillers, R., Jeansoulin, R. (eds) (2006). Fundamental of Spatial Data Quality. London: ISTE.

Devillers, R., Stein, A., Bedard, Y., Chrisman, N., Fisher, P., Shi, W. (2010). Thirty years of research on spatial data quality: achievements, failure and opportunities. Transaction in GIS, 14 (4), 387–400. DOI: http://dx.doi.org/10.1111/j.1467-9671.2010.01212.x

EC. (2007). 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).

Frank, A. U., Grun, E., Vasseur, B. (2004). Procedure to select the best dataset for a task. GIScience 2004, LNCS 3234, 81–93. DOI: http://dx.doi.org/10.1007/978-3-540-30231-5_6

Geodetic and Cartographic Law (2014). Journal of Laws No. 2014 entry 897.

Goodchild, M. F. (2008). Spatial accuracy 2.0. In J. Zhang and M. Goodchild (Eds), Spatial Uncertainty: Proceedings of the Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 1,1–7. Liverpool: World Academic Union.

ISO 19 115:2014 Geographic information - Metadata - Part 1: Fundamentals. Geneva, Switzerland: ISO/TC211.

ISO 19157:2013 geographic information: Data quality. Geneva, Switzerland: ISO/TC211.

ISO/TS 19103:2005 Geographic information - Conceptual schema language. Geneva, Switzerland: ISO/TC211.

Ivánová, I., Morales, J., de By, R. A., Beshe, T. S., Gebresilassie, M. A. (2013). Searching for spatial data resources by fitness for use. Journal of Spatial Sciences, 58 (1), 15–28. DOI: http://dx.doi.org/10.1080/14498596.2012.759087

Juran, J. M., Bingham, R. S. (1974). Service industries. In J. Juran, F. Gryna. Jr., & R. Bingham (Eds.), Quality control handbook (47-1, 47-35). New York: McGraw-Hill.

Juran, J. M. (2010). Juran's quality handbook. New York [u.a.]: McGraw Hill. 6th edition.

Kahn, B. K., Strong, D. M., Wang, R. Y. (2002). Information Quality Benchmarks: Product and Service Performance. Communications of the ACM, 45(4ve), 184–192. DOI: http://dx.doi.org/10.1145/505999.506007

Li, G., Fang, Ch., Pang, B. (2014). Quantitative measuring and influencing mechanism of urban and rural land intensive use in China. Journal of Geographical Sciences, 24 (5), 858–874. DOI: http://dx.doi.org/10.1007/s11442-014-1125-z

Lush, V., Bastin, L., Lumsden, J. (2012). Geospatial data quality indicators. In C. Vieira, V. Bogorny, A. R. Aquino (Eds.), Proceedings of the 10th international symposium on spatial accuracy assessment in natural resources and environmental sciences. International Spatial Accuracy Research Association (pp. 121-126), 10th international symposium on spatial accuracy assessment in natural resources and environmental sciences. Brazil, Florianópolis, 10–13 July.

Montero, J-M., Chasco, C., Larraz, B. (2010). Building an environmental quality index for a big city: a spatial interpolation approach combined with a distance indicator. Journal of Geographical Systems 12(4), 435-459. DOI: http://dx.doi.org/10.1007/s10109-010-0108-6

MAiC. (2013). Regulation of the Minister of Administration and Digitization of 5 September 2013 on the organization and the mode of conducting state geodetic and cartographic (in Polish). Journal of Laws 2013, entry 1183.

Nebert, D.D. (ed) (2004). GSDI Cookbook, Version 2.0, 25 January 2004 http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf

Paradis, J., Beard, M. K. (1994). Visualization of Spatial Data Quality for the Decision Maker: A Data Quality Filter. URISA Journal, 6 (2), 25–34.

Rau, J. N., Chen, L. C. (2003). Robust reconstruction for building models from three-dimensional line segments. Photogrammetric Engineering and Remote Sensing, 69 (2), 181–188. DOI: http://dx.doi.org/10.14358/pers.69.2.181

Redman, T. C. (2000). Data quality: the field guide. USA: Digital Pr.

Steiniger, S., Lange, T., Burghardt, D., Weibel, R. (2008). An Approach for the Classification of Urban Building Structures Based on Discriminant Analysis Techniques, Transactions in GIS, 12 (1), 31–59. DOI: http://dx.doi.org/10.1111/j.1467-9671.2008.01085.x

Surveyor General of Poland. (2003). Wytyczne techniczne, Baza danych Topograficznych – wersja 1, Główny Geodeta Kraju (Technical Guidelines, Topographic Database – version 1), the Surveyor General of Poland.

Takashima, M., Hayashi, H., Nagata, S. (2003). Monitoring spatial distribution of population and buildings using DMSP night-time imagery and its application for earthquake damage assessment. IEEE International, 4, 2430–2432, DOI: http://dx.doi.org/10.1109/IGARSS.2003.1294465

Wright, E. J. (2006). Fitness for use – to support military decision making. In M. Caetano, M. Painho (Eds.), 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (760-769). 5-7 July 2006, Lisbon, Portugal, ISARA (International Symposia on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences).

Zhang, X., Ai, T., Stoter, J. (2010). Characterization and detection of building patterns in cartographic data: two algorithms. In: Joint international conference on theory, data handling and modelling in geospatial information science (SDH' 2010), XXXVIII (2), 261–266.

Zhang, X., Ai, T., Stoter, J., Kraak M-J., Molenaar, M. (2013). Building pattern recognition in topographic data: examples on collinear and curvilinear alignments. GeoInformatica, 17 (1),1–33. DOI: http://dx.doi.org/10.1007/s10707-011-0146-3

Bielecka E. (2015). Geographical data sets fitness of use evaluation . Geodetski vestnik, 59 (2): 335-348.
DOI: 10.15292/geodetski-vestnik.2015.02.335-348

Assoc. Prof. Elżbieta Bielecka, PhD
Military University of Technology, Faculty of Civil Engineering and Geodesy
gen. S. Kaliskiego 2
00-908 Warsaw, Poland
e-mail: ebielecka@wat.edu.pl