How Laws affect Data Quality

Gerhard Navratil
TU Vienna, Institute of Geoinformation and Cartography

Abstract. When dealing with data quality we usually think about geometric or semantic quality. These types of quality emerge from the method of data collection. Legal rules have an influence on both, geometry and semantics. Laws also impose limitations on the data and we usually ignore these limitations. The limitations may have a higher influence than technical aspects. In this paper I discuss the different types of limitations laws can impose on data and the effects on data quality. The paper also shows a way to deal with legal influences on data quality.

1. LEGAL RULES DEFINE SEMANTICS IN WORDS

Data is the most expensive part in information systems. According to Frank the cost of hardware, software, and data are in the proportions 1 : 10 : 100 (Frank 1995). Sharing data can save significant amounts of money but requires a description of the data set's quality. Then the user can check if the data set is suitable for his task.

Laws and decrees stipulate data quality. A simple example is topographic maps. In Austria the official topographic maps are updated every 7 years. The temporal quality of a new map is therefore between a few weeks (the time necessary for processing and inserting gathered data and for printing the map) and 7 years.

Laws and decrees also create the framework for using data. Boundaries for pieces of land, for example, are a line on a map or a set of coordinate values in a database. The semantics of the line derives from the legal status of the line. Different countries usually have different laws resulting in different status of a piece of data that seems to be identical. Thus we need to know the laws concerning the data if we want to grasp the semantics of data.

Current focus of research is on measurement of data quality rather than on semantic operability. In this paper I want to combine semantics with the legal rules creating it. Kuhn's semantic reference system (Kuhn and Raubal 2003) provides a framework for the integration of the legal semantics.

The combination of data sets from different sources usually includes technical and semantic problems. Technical differences can usually be overcome because there are mathematical solutions for problems like different coordinate systems. Semantic problems are more difficult to solve because they are more likely to be missed. Kuhn's example of the different semantics of ferries and bridges relates to a real incident where a navigation system did not include the semantics (Observer 1998). Problems also occur if a navigation system for cars does not distinguish between bridges for cars and pedestrians. Such errors emerge from misunderstanding

semantics during processing. The likelihood of such errors increases with the number of data sets with different semantics.

As an example throughout the paper I will use the Austrian cadastral system. The Austrian cadastre has been originally introduced to collect land tax (Twaroch and Muggenhuber 1997). The parameters needed for this task are

- the size of the piece of land owned by a person and
- the way the land is used (determines the maximum income for the owner).

Other aspects like the shape or position of the piece of land are less relevant for tax stipulation. In Austria, the shape was used only to determine the size and to provide a map showing the spatial relation between the different pieces of land.

Differences in legal definitions provide more difficult problems. How can we deal problems produced by the different status of a boundary line in a cadastral map in different countries? Even in Austria there are two different types: The line in the new, coordinate-based cadastre ('Grenzkataster') is fixed by coordinates and defines the boundary. The line in the old fiscal cadastre is only an indication for the position of the boundary and the real definition is based on agreement between the owners of the bounding pieces of land.

2. DATA QUALITY

Why are data quality descriptions necessary? Descriptions of data quality are only necessary if data collector and data user are not the same. The necessary quality is known if producer and user are the same person or at least closely related in one agency. This implies communication of data quality. Description of data quality is necessary if there is no direct link between producer and user, i.e., if the user cannot influence the quality of the data produced. In this case the user needs a detailed description of the data quality to select a suitable data set for his task. It may even be necessary to modify the user's task to adjust the task to the data available.

Data sharing results in the separation between producer and user of data. In a traditional data production environment a user asks a data producer to collect data using a specific method. The result is a data set fit for the user's task. Unfortunately production costs for data are usually high whereas copy costs are almost zero. Using an already existing data set will therefore reduce the costs significantly. On the other hand, selling a data set to more than one user will increase the profit for the producer or will allow him to sell the data much cheaper than otherwise. This will lead to a situation where the data set was not intended for the task it is used for. A description of the data quality then provides two benefits:

- The user can check if the data set is adequate for his task. Objective quality measures could even be used by search engines to automatically search the Internet for suitable data sets.
- The producer is not liable for misuse of his data sets. Quality measures do not prohibit misuse of spatial data. In some cases it will even be necessary to use inadequate data since no other data is available and time pressure prevents data collection. Experienced users will know the limitations imposed by the inadequate

data and will be cautious when drawing conclusions. Amateurs might not be able to do so. In that case the producer is not liable for damage resulting from deviations in the data set matching the stated quality measure of the data (Frank 1998).

Data quality consists of a set of quality measures. In general data quality is necessary for attributes, topological properties, positional and temporal information (Laurini and Thompson 1992, p. 104). Recently there have been approaches to manage data quality within GIS from researchers as well as from software vendors (Qiu and Hunter 1999; ESRI 2002).

3. HOW LAWS INFLUENCE DATA QUALITY

Laws affect the data available in a variety of ways. The following list contains two of the most important aspects.

- Quality of the data collection: Laws often influences directly the quality of the data collection by stipulation of limits for quality parameters.
- Purpose of the data collection: Data collected in fulfillment of laws has a specific purpose. Cadastres provide data for land tax, registration of residence allows authorities to contact specific persons, and air quality measurements try to detect illegal air pollution. Each of these tasks requires specific data. Some aspects of data quality may also be connected to the purpose of the data collection. Air pollution data for providing a warning on dangerous air pollution cannot work with an update interval of a decade. Thus the purpose of the data collection does not only define the attributes but also affects data quality.

Access restrictions are another aspect of data quality if only limited quality is available to the public due to data protection laws or similar regulations. Data protection laws grant the right of the person for privacy of specific types of data. This shall avoid situations where a person suffers disadvantages due to available data, for example medical records. Some organizations will need this type of data but the data should not be generally available. The two main reasons which may lead to such restrictions are:

- Restriction to protect the personal data of citizens. Personal data like an HIV-infection could probably be used against citizens. The aim of data protection initiatives is to avoid such misuse of data. The simplest method to do this is aggregation. Different user groups may even have different granularity. Another method is prohibiting specific queries. The Austria land register, for example prohibits the search for all pieces of land a specific person owns.
- Restrictions to keep data secret due to political or military reasons. The military often tries to cover up locations with important installations to make emergency planning more difficult in case of war. Even with satellites providing data maps are still falsified in some areas. This leads to wrong quality assessments by the user. Without field inspection he may not be able to detect falsifications.
  In both cases quality as well as availability depends on legal definitions.

3.1 Laws Defining the Quality of the Data Set

In many cases laws specify the quality for data capture. The decree for surveying (1994) stipulates the standard deviations for points in the coordinate-based cadastre (§7). The maximum standard deviation allowed is 15cm for the point, which results in 10cm standard deviation for the coordinates. This definition influences the method for data collection as well as the quality of the resulting data set. Since users only get the coordinates without additional information in regard to the measurement of the points, this definition is the only quality measure users have access to.

Another important data source usually is statistical data. Collection of statistical data in many countries is done by governmental organizations. The rules for the data collection is written down in laws. These rules may contain methods to granting data protection like merging answers for all persons living in a building or area. There may also be rules prohibiting questions on a specific topic like diseases and sexual or political preferences. The first set of rules restricts the quality of the data whereas the later rule restricts the availability of data.

Sometimes laws can even contradict each other with quality assessments. The decree for surveying not only specifies the standard deviation for coordinates but also the precision for area measures (§10 (2)). The size of an area must be given in square meters. A simple analysis shows that these quality measures do not match (Navratil 2003) since a standard deviation of 10cm for coordinates leads to a standard deviation of up to 8% of the area (typically more than 1m2) when dealing with a cadastral data set.

3.2 Laws Defining the Purpose of the Data Set

The purpose of a data set reflects in the observations performed when producing the data set. The Austrian cadastre provides a good example. The original purpose of the cadastre was fair taxation. This leads to a solution with high quality in the agricultural areas and low quality in the residential areas since residential areas produced no income and tax was based on maximum possible income. The parameters necessary for taxation were size of the area and use of the land. Thus these parameters were observed and stored as attributes for each piece of land. Due to limitations of the methods used the quality of the area was approximately 10%. Since taxation before the invention of the cadastre was done more or less arbitrarily this was still a major improvement in justice.

During the 20th century the usage of the cadastral maps changed since they were the only large-scale maps (scales of 1:500 to 1:5.760) available for the whole country. The demand for quality improved dramatically and additional attributes were necessary. However, it is almost impossible to improve a data set like a cadastre without complete reconstruction. Complete reconstruction of the cadastre would incur a productive cost.

In the 1960's the idea developed that a cadastre could also serve as proof for boundaries of land. This idea required changes in the cadastral law and finally led to a system where each boundary point has coordinates in a national reference frame, the coordinate-based cadastre (1968). The coordinates allow reconstruction of the

boundary points in the field. Legal security also demands that these coordinates determine the boundary in the new system whereas the agreement of the owners and evidence in the field defined the boundary in the old system. Therefore the boundary lines in the old and the new system have different effects.

Both cadastral systems still run in parallel, but a piece of land can only be registered in one of the systems. The legal differences between the systems, however, are critical. Some operations are possible in both systems, like

- computation of area or
- computation of land tax.

Other operations are only possible with limitations, like

- subdivision of parcels (newly created parcels are automatically part of the coordinate-based cadastre) or
- merging of parcels (all parcels concerned must be part of the same cadastral system).

Other operations are only possible in one of the systems, like

- definition of boundary in discussion with the owners of the land (only possible in the old system) or
- marking the boundary in reality without the possibility of objection by the owner of land (only possible in the coordinate-based cadastre).

In addition to these changes in the operations the representation changed. Each boundary point in the old cadastral system was represented as a circle on the cadastral map. In the coordinate-based cadastre the points are represented by coordinates. During the creation of the digital cadastral map in the 1990's also the boundary points in the old cadastral system got coordinates. However, these coordinates have the same legal status as the circles on the maps had. Thus the representation of boundary points belonging to the old and the new system is equal. The only difference is a flag stating if the point is from the coordinate-based cadastre or not.

Computer programs dealing with cadastral maps from Austria should not ignore that flag. The quality of coordinates from the old cadastre is very low compared to the quality of the coordinates in the coordinate-based cadastre. Many computations are based on coordinates like computing the size of a piece of land. Such computations should only be applied to points from the coordinate-based cadastre since there is no quality measure for the other type of coordinates (the boundary is not based on the coordinates but on the real situation). If the points are not from the coordinate-based cadastre the size of the piece of land should be extracted from the list of land sizes introduced for taxation because these numbers are the official numbers.

4. CONNECTING LEGAL ISSUES TO DATA

Kuhn proposes semantic reference systems to store the semantics of data. Semantic reference can be seen as wrapping the data. The wrapper has a defined interface and only accepts queries permitted by the interface. The wrapper defines the process methods needed to access the data. When computing areas for a cadastre, for example, the wrapper must decide if the area can be computed from coordinates or if the area must be extracted from a list of areas.

A semantic reference system dealing with legal issues requires a method to formalize these issues. It is possible to create formal models for laws (Navratil 2002; Navratil 2003). The method creates an axiom-based algebraic description of the legal definitions. It should be possible to apply the method to legal definitions affecting data quality. This will provide a tool to directly link the legal situation to the data. This provides the user of the data with an unambiguous description of the legal aspects of the data. Since mathematics serves as the language used the model is even independent of the language the law is written in. Minor dependencies may emerge from necessary function names but these limitations are small in comparison to the problems of reading law texts in a foreign language. Assuming a semantic reference system as the basic structure the axioms derived from the laws define the functionality of the semantic reference system. The functionality finally represents the limitations imposed by legal rules.

A simple example shall clarify the approach. Access restrictions like those imposed by data protection laws require different user interfaces for different user groups. As said before the Austrian land registration law prohibits queries based on the name to get a list of all pieces of land a specific person owns. Some user groups, however, must have access to such a list: The financial authority to compute the land tax for a person or courts to determine the financial situation of a person in case of inheritance or claims against that person. The user interface requires a parameter specifying the type of user to determine the access rights. In some cases the data will not be provided, in other cases the quality of the data received depends on the status of the user, and if no access restrictions are in place the user will receive all data available for his query.

4.1 Modeling the Quality of the Data Set

Merging data sets is similar to the search for the least common denominator. Corresponding properties need to be matched and inconsistent properties must be dropped. Quality assessments can only be set to the lowest level achieved by all data sets. An example for such a merge is support for agricultural land by the European Union. The simplest way to determine the areas is looking in the cadastre. Unfortunately the members of the EU have different cadastral systems with different quality for the area estimation. Merging these systems would assume the quality of the system with the lowest quality for all areas. If a member does not have a complete and consistent cadastre the result would be completely useless. Therefore, the EU only defined quality measures and did not specify how to obtain that quality. Thus in principle they did the same on an international level as the member states did on the national level, they specified the quality for the result.

Modeling the quality of the data set can be done in different ways. The easiest method is providing a unique value for the whole data set, like the Austrian cadastre does for the coordinates of the boundary points. A measurement-based system allows a more sophisticated method by applying a standard deviation to each point in the data set even providing information on the correlations between the points.

4.2 Modeling the Purpose of the Data Set

The purpose of the data set has an influence on the data capture and therefore an influence on the data collected. Thus the purpose influences the categories available. It should also reflect the possible use of the data set. This could be done by providing a function 'usable_for' which takes the intended use as a parameter and provides a measure for the usability of the data set for this type of use.

Another solution would be that the user interface only allows functions which are meaningful for the data set. It is, for example, not useful to compute slopes from many European cadastral data sets because they do not contain height information. This could be done, however when using topographic maps as data sets. Trying to get information on the use of land from topographic maps is similarly useless.

5. CONCLUSIONS

It became clear that the influence of legal definitions on data quality is important. Three different types of influences became visible: Intention, quality assessment, and security measures. Each of these influences has an impact on the way we have to treat the data set. Especially when merging data sets of the same type (for example cadastral data sets from different countries) we have to take into account the legal situation.

Semantic reference systems can provide a frame to store semantic reference and formalizations of the legal definition lead to the functions necessary. This allows using the data in a way that prohibits misuse as far as possible. Users would get warnings if the intended use does not fit to the type of data. Moreover, when merging data sets the semantic reference system can provide the transformation between the reference systems.

The benefits of legal aspects in semantic reference systems are manifold. The main advantage is the restricted risk for the inexperienced user. The defined user interface limits the number of operations applicable to the data set. This reduces the risk that a user applies an operation that is not suitable for the data set. On the other hand the experienced user gets some information on the quality aspects of data sets. Using data sets from different countries always includes the risk of unexpected legal restrictions. A semantic reference system will clearly show legal peculiarities and will therefore reduce the risk. Both errors typically occur with automated procedures, too. Computer programs can be redesigned to use the simple interface of the semantic reference system to avoid problems with inappropriate data sets. Thus experienced as well as inexperienced user will benefit and legal definitions in semantic reference systems will even improve the efficiency of automated processes.

It became clearly visible that legal rules specify semantics of data. This also must have an effect on new legal rules. Laws defining slightly different semantics for data necessary force the authorities to collect data even if other authorities may have similar data. Transformation between different semantics includes loss of information and the result may not be accurate enough. Future research should investigate the

connection between slightly different semantics in laws and the data sets collected to fulfill the laws. Maybe this can then help avoiding multiple collection of data.

ACKNOWLEDGEMENTS

This work has been supported by the project ReviGIS sponsored by the European Commission.

REFERENCES

(1968). Bundesgesetz über die Landesvermessung und den Grenzkataster. BGBl. Nr. 124/1969.
(1994). Verordnung des Bundesministers für Bauten und Technik über Vermessungen und Pläne. BGBl. Nr. 562/1994.
ESRI (2002). Metadata and GIS. Redlands, CA, ESRI White Paper.
Frank, A. U. (1995). The Economics of Geographic Information. Geographic Information Systems - Materials for a Post-Graduate Course. A. U. Frank. Vienna, Department of Geoinformation, TU Vienna. 3: 745-801.
Frank, A. U. (1998). Metamodels for Data Quality Description. Data quality in Geographic Information - From Error to Uncertainty. R. Jeansoulin and M. Goodchild. Paris, Editions Hermès: 15-29.
Kuhn, W. and M. Raubal (2003). Implementing Semantic Reference Systems. AGILE, Lyon, France, Presses Polytechniques et Universitaires Romandes.
Laurini, R. and D. Thompson (1992). Fundamentals of Spatial Information Systems. San Diego, Academic Press.
Navratil, G. (2002). Formalisierung von Gesetzen. Vienna, Institute for Geoinformation, Vienna University of Technology.
Navratil, G. (2003). Modeling Processes defined by Law. AGILE, Lyon, Presses Polytechniques et Universitaires Romandes.
Navratil, G. (2003). Precision of Area Computation. ESRI 2003 - 18. European User Conference, Innsbruck, Austria.
Observer (1998). BMW computer makes driver turn into drinker.
Qiu, J. and G. J. Hunter (1999). Managing Data Quality Information. International Symposium on Spatial Data Quality, Hong Kong, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University.
Twaroch, C. and G. Muggenhuber (1997). Evolution of Land Registration and Cadastre - Case Study: Austria. Joint European Conference on geographical information.