Quality of Spatial Data for e-Government from an Ontological View

Gerhard NAVRATIL, Austria
Andrew U. FRANK, Austria

Key words: data quality, ontology, e-government.

SUMMARY

Discussion on quality of spatial data is traditionally separated in different groups. Data quality mainly covers the observation process and the process of data manipulation. Uncertainty describes the problems of inaccurate boundaries between different classes. Usability concentrates on the view of the data user and tries to answer the question if a data set is usable for answering a specific question or not. Unfortunately the members of each group do not pay enough attention to the developments of the other groups (Fisher 2003).

Previous investigations showed that there are impacts of legal definitions on the quality of available data (Navratil 2004). In addition socially constructed objects are used differently than objects constructed from observing reality. The process of defining the boundary of a land parcel differs from the process of defining the extent of a lake.

The paper investigates the different types of decision processes and shows that processes of e-government must deal with a different type of quality than technological processes. Using the 5-tier ontology we can separate different levels and show how the different aspects of quality emerge and how they influence decisions.

1. INTRODUCTION

E-government is a priority of the Austrian EU presidency. The goal is to create cross-border services that provide standardization for administration (Federal Chancellery 2006). This requires efforts to harmonize processes and legal regulations.

Working e-government requires, among other things like Internet access for a large part of the society, predictability. One of the main advantages of e-government is that these processes are user-friendly since applications are not restricted to office hours. This “24/7” access (i.e., 24 hours a day, 7 days a week) is also one of the major benefits of e-government for public participation processes (Carver 2001). In addition to the eliminated temporal restriction, there is a gain of time for the user. These benefits are only valuable if the outcome of an administrative process is predictable. Digital access to a land register, for example, simplifies the process of getting data on the legal situation of land. The possibility to send applications for registration in the land register in digital form eliminates the necessity to travel to the land register. These benefits will vanish if the data in the land register is not complete. The outcome of an application is unpredictable if the person named

as owner in the online access may be wrong. In this case the application cannot succeed and the time spent on the application will be lost. The users will fall back upon the traditional methods to avoid such losses if the risk for failure is high.

In this paper we discuss the quality issues of predictability. We use an ontological approach to identify the key elements for predictability. This requires a discussion of different approaches to quality and a model for decision making in the legal realm.

The paper is structured as follows: Section 2 introduces the 5-tier ontology, which allows separation of different aspects of reality and allows separation of technical from legal problems. Section 3 then discusses quality from different viewpoints. In section 4 we deal with decision processes separating decision processes in technical systems from those in the legal realm. In section 5 we discuss the consequences of the separation for e-government. Finally, section 6 contains some conclusions.

2. THE 5-TIER ONTOLOGY

Ontologies are a subject where a large amount of publications exists. Agarwal discussed the different viewpoints and objectives of the ontology discussion (Agarwal 2005). He separates a philosophical approach leading to top-level ontologies and a software development/data modeling approach leading to domain specific ontologies.

The 5-tier ontology (Frank 2001) is a top-level ontology separating different levels of geographic reality, both physical and social. Frank starts from the separation between the physical reality and the results of observing this reality, a discussion that goes back to the ancient Greek philosophers. Additionally, Frank introduces layers for objects with properties, socially constructed objects, and the subjective knowledge. In the following, land and the different views on land will be used as an example.

Tier 0 describes the physical environment we live in. The underlying assumption is that there is only one single physical environment. Tier 1 contains the results of observing tier 0. The separation of these two levels dates back to the Greek philosopher Plato. Plato pointed out the necessity to separate reality from our knowledge of it. Frank assumes that each point in space and time has determined properties and that space and time are the fundamental dimensions of this reality. The observation allows us to determine the surface of the earth, a step necessary to define land.

Tier 2 groups points into objects. The observations reveal properties of the points and sets of points with unique properties are considered as objects if they continue in time. Temporal constructs for objects have been defined by Al-Taha and Barrera (1994), extended by Hornsby and Egenhofer (1997), and formalized by Medak (2001). These are real world objects because observation of reality allows separating them. For example, the object land is defined by its separation from air and water. This makes islands and continents objects. However, these areas of land are subdivided further into objects, e.g., by placing fences or walls or by splitting the object land along natural dividers like rivers or ridges.

Tier 3 contains objects separated from each other by social constructs. Socially constructed reality (Searle 1995) assigns special meaning to real world objects if they are created using special social processes. The creation of a land parcel, for example, must follow land registration procedures. The result is a piece of land with the socially defined attribute ownership. Simply placing a fence is not considered sufficient in many jurisdictions.

Finally, tier 4 is the subjective view of cognitive agents. Each agent makes different observations. This influences his beliefs about the world and influences his decisions. Value of land may, for example, vary with the person judging the value. Land will usually be more valuable for people who lived their whole life on that land than for strangers.

3. MEASURES OF QUALITY

The necessity for measures of quality emerges from two sources. Firstly, the observations of reality in tier 1 are imperfect and will contain deviations from the 'real' values. Secondly, the rules used to form the objects in tier 2 may be fuzzy. The research fields are data quality or uncertainty depending on the major source.

3.1 Data Quality

Originally, quality expressed the superiority of a manufactured good or a high degree of craftsmanship or artistry. Quality is the degree of excellence of a product, service, or performance. Within manufacturing, quality is achieved by management and control of the production process. Many of these issues apply to the quality of data, since data are the result of a production process called observation. The reliability of the process imparts value and utility to the database.

It is well established that the results of observation processes contain errors. The physical reality in tier 0 is error-free and thus there is a predefined value, which should be returned by an observation process. However, the result v of observation process deviates from this 'real' value v' by the value ε:
$v'+\varepsilon = v$.

Data quality measures must describe the deviations. Measures to describe data quality are separated in different aspects (Guptill and Morrison 1995; Wang and Strong 1996; Veregin 1999):

- **Lineage** describes the methods used to obtain the data and the processes applied to the data.
- **Accuracy** describes the quantity of the variations in the observation process. It is applied to both, the position and the values of the attributes.
- **Completeness** is the relationship between the occurrences of the phenomenon represented in a data set and the abstract universe of all occurrences of the phenomenon. Completeness consists of data and model completeness. Data completeness is the number of missing elements that should be in the data set

(omission). Model completeness refers to the agreement between the database specification and the abstract universe (Brassel, Bucher et al. 1995).

- **Logical consistency** reports the results of checks on logical contradictions. Parcels of cadastral data sets, for example, should not overlap and water should not flow upwards.
- **Semantic accuracy** deals with cases where the phenomenon represented by a specific class in the data set does not fulfil all requirements for this class (Salgé 1995). For example, woodlands stored as oak woodlands may not match the definition for oak woodlands on which the database is based.
- **Temporal accuracy** or **currency** is an indication how up-to-date the data set is.

Examples for data sets where data quality is useful and easy to provide are digital terrain models or satellite images. However, as soon as the observations must be classified, e.g., when identifying mountains or rivers, discussion of data quality is not sufficient. Results from the uncertainty discussion become relevant for the classification result. Uncertainty describes the problems with the class whereas data quality describes the quality of the boundary definition and the attribute values based on the quality of the observation process.

3.2 Uncertainty

Classification of objects is based on the concept used to think about space. Discussion of concepts leads to uncertainty measures if the concepts do not have crisp boundaries. Unfortunately, this is true for most of the concepts in geography. We cannot specify, for example, how high a protuberance must be to be called a mountain. It is also difficult to specify the number of trees necessary to form a forest. Fisher separates four main aspects of uncertainty (Fisher 1999; Fisher 2003):

- **Errors** emerge from wrong observations. In contrast to accuracy in data quality here we do not assume normal distribution for the observations.
- **Vagueness** is based on the concept of fuzzy set theory as introduced by Zadeh (1965). A classification may result in an ambiguous situation if based on vague concepts. Classification of spatial objects may not be possible unambiguously. A protuberance, for example, may fit the class definitions of both, mountains and hills.
- **Ambiguity** arises if a classification produces different results if using different procedures.
- **Discords** are contradictions between data sets that are based on different classification schemes. In one set a protuberance may be a mountain and in the other one it may be a hill.

Some aspects of uncertainty are only important in geographic space and do not exist in the socially constructed reality. Data sets containing geographical features must cope with the problem of vagueness since there are no sharp boundaries between the concepts. In the socially constructed reality, however, there are strict, non-overlaping definitions written down in laws. Forrest, for example, is defined in the Austrian law as an area used for forrestry (Austrian National Assembly 1975). This eliminates all problems of tree size, density, and quantity. There is also a defined strategy if problems arise during the classification process. The case is presented to a judge and a decision is made. This process

solves problems of vagueness and at the same time problems of ambiguity since the classification is only done once and the result is correct by definition.

Discord emerges if using different classification schemes and comparing the results. Legal systems are defined in a way that each word has exactly one definition. Forest, for example, has no second definition in the Austrian law.

4. DECISION PROCESSES

Data are not collected without reason. Data are used to make decisions. These decisions may have an impact on the structure of tier 2 or tier 3. The decision if a building moves may have an impact on further actions. Observations showed, for example, that the Leaning Tower of Pisa tilted due to the pressure of the building on the soft soil. Several attempts were made during the second half of the 20th century to prevent the tower from toppling. The results of these attempts were observed and in 2001 the tower has been declared stable (Wikipedia 2006). This is a typical example for a technical decision process. Decisions in legal systems work differently as there is no feedback loop of observations. This section discusses decision processes firstly in technical and secondly in legal systems to show the similarities and differences between these kinds of systems.

4.1 Decisions in technical systems

The decision in a technical system is based on observations. Since the observations contain deviations, the parameters derived from the observations will also contain deviations. These parameters are then used to make decisions (Navratil and Frank 2006, submitted). The example of the Leaning Tower of Pisa shall clarify the process.

The detection of a movement requires observations of the position of a specified point. In the case of the Leaning Tower of Pisa only the height is of interest. The time period between the observations is selected such that a statistical test can be used to determine the existence of a movement. Each observation has a specified accuracy. Thus a repeated observation at the same time will produce a slightly different result. Statistically speaking the results of an observation process are the results of a randomized experiment. The result of the observation processes is thus the height h_1 with standard deviation σ_1 at the time t_1 and the height h_2 with standard deviation σ_2 at the time t_2. The statistical test must then determine if the difference h_2-h_1 between the heights can be explained by the standard deviations of the heights or as a result of a movement. In this simple case the test checks the hypothesis if the expected values for both observations is the same or not (Reißmann 1976: 323 ff). A general method for such tests is deformation analysis (Niemeier 1985).

The result of a statistical test is based on the confidence level used. Statistical tests always include the chance of an error. Decisions are made based on the result. The decision may be reconsidered if the statistical test shows a different result when using a different confidence level. New observations of the same elements may also require changing the decision. The decision that the Leaning Tower of Pisa is stable may be reconsidered, for example, if new observations suggest a movement of the tower.

The advantage of decisions in technical systems is that a logical chain can be optimized. The result of such an optimization can take the quality of the observations into consideration. The problem with this approach is that humans cannot produce the results without tools like mathematics. A human can, for example, approximate the centre of gravity for a homogeneous 2D-object. This is not possible for inhomogeneous objects. In this case mathematical approaches are necessary. Complex chains in technical systems contain too many variables to be handled by humans and thus the result of an optimization process is not easy to check.

4.2 Decisions in the legal realm

Legal decisions are made in a different way. The basic concept of legal decision making is the subsumption. Complex decisions are subdivided into a series of small and simple decisions and only after deciding on one step the next step is discussed. Law is constructed by regulations consisting of conditions and consequences. During a legal process the situation is compared with the conditions. In many cases there is room for decisions, which can be modeled with a fuzzy membership function as shown in Figure 1. However, the decision itself must be binary. The situation either matches the conditions or it does not. Otherwise subsumption would not work.

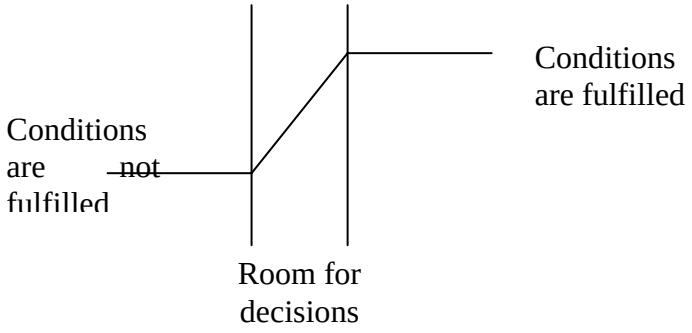

Figure 8: Legal matching process

Murder, for example, is the “unlawful killing of a human being with malice and forethought” (Black 1996). During a trial the prosecutor must show that all conditions are fulfilled whereas the advocate tries to show that at least one point is in doubt. The consequences are decided only if the court decides that all conditions for murder are fulfilled. The discussion of consequences does not include the question if the action was murder since this was already answered.

Difficulties arise if the legal decision process uses values derived by a technical decision process. For a judge a measured value of 3.96m is smaller than 4.00m whereas a statistical test may show that the values may have no significant difference. This contradiction arises from the legal inability to deal with tolerances that are normal in technical systems

(Twaroch 2005). This problem has been discussed in the context of the Austrian law for surveying by Peters (Peters 1974: p. 39).

Subsumption transforms the real situation to a socially constructed situation. The real situation is part of tier 2. The elements of this situation are real actions and objects, which are then compared with the definitions of the socially constructed actions and object. An action is murder if the court decides that all conditions for murder are fulfilled and a measured distance is larger than 4.00m if the judge decides that the distance is larger. The results are then no longer real objects but members of a socially constructed class. Thus the results of the subsumption process belong to tier 3. As members of tier 3 the correctness of these objects is out of question (unless decided otherwise after an appeal).

The advantage of the legal decision process in contrast with the technical decision process is the simplicity of the decisions. As stated above, a chain of decision in the legal realm consists of a number of simple, independent decisions that can be treated by humans. Contrary to the technical system the result may not be optimal with respect to an overall criterion but each decision in the chain can be made by a human.

5. CONSEQUENCES FOR E-GOVERNMENT

E-government shall simplify administrative processes for the users. Austria has transformed several processes from the traditional form to the electronic form. These processes include (Federal Chancellery 2006):

- application for tax declaration,
- application for inscription in the land register,
- application for support for agricultural areas, and
- publication of laws.

These processes have two characteristics in common. Firstly, the results of the processes are predictable and secondly, the objects and figures necessary for the processes belong to tier 3. The web page for tax declaration, for example, has a number of different categories like income, extraordinary financial burdens, or tax free allowances. These categories are defined in the tax law. The user fills in his numbers and the web page can predict the result of the application assuming that all numbers are accepted. The prediction can only fail if the user cheats and the tax authority notices the attempted fraud. The numbers filled in by the user belong to tier 3. The social context of the user influences the numbers. Journalists or politicians, for example, are allowed to assert a fixed amount of advertisement costs whereas other social groups may have no such fixed amount. The costs are not connected to the real bills paid for advertisement, which would be objects from tier 2. Even if the user must be able to provide bills, the figures are no objects from tier 2 since the difference between a bill for advertisement costs and a bill for other costs is the legal definition of advertisement costs. This definition, and thus the bill itself, is part of tier 3.

Predictability requires that the data accessible to the user are the same as the data used to determine the result of the process. An application for inscription in the land register provides an example as already briefly mentioned in section 1. The land register lists the

owner of land and shows the encumbrances and restrictions imposed on the land. Registration is allowed if the owner of the land agrees. The process of inscription becomes unpredictable for the user if the owners visible to the public are incomplete or outdated. Let us assume that land is sold to a person. There will be a time period until this change is visible to the public. Problems may arise if this time period is too long since persons may try to sell land twice. This can destroy the trust in the electronic process because the second buyer will not become owner. The result of the missing trust may be that the public returns to the traditional, analogue processes and e-government is not used.

The example showed that currency, attribute accuracy, and completeness of necessary data are crucial for the predictability. Logical consistency for electronic processes is equally important as for the traditional processes. Problems of uncertainty may arise if the laws are ambiguous and the result of processes is not completely determined by data and legal rules.

Guriev showed that there is a connection between the complexity of administrative systems and the amount of corruption (Guriev 2004). Corruption may cause the failure of e-government if the level of corruption is too high. Processes of e-government shall disconnect the clerks from the applicants thus eliminating possible bribes. However, bribes are usually paid to reduce the time until completion of the process or to ensure a positive result. Users will find methods to still pay bribes if the electronic processes are too slow or the results do not meet their demands. Thus e-Government is no tool to eliminate corruption.

6. CONCLUSIONS

We have seen that the prime concern for e-government is predictability. It is necessary that the result of an electronic process is based on a rational and thus predictable decision making process. One of the key elements to predictability is clear legal concepts of the elements involved in the process of decision making. The clarity of legal concepts eliminates problems of uncertainty. Another key element is the ’correctness’ of the data used in the decision making process. If the user and the administration use different data, the result of the process will be unpredictable for the user. Therefore the data accessible for the user must be the same as the data used by the administration. This leads to high quality demand on currency, attribute accuracy, and completeness.

We must see, however, that e-government is not solution for all problems of administration. The prime goal of e-government is simplification for the users by eliminating the restriction to office hours. This works for well established processes as the examples from the Austrian administration show. However, problems like corruption will not be solved by e-government because it is not possible to remove the traditional form of application.

REFERENCES

Agarwal, P. (2005). "Ontological Considerations in GIScience." International Journal of Geographic Information Science **19**(5): 501-536.

Al-Taha, K. and R. Barrera (1994). Identities through Time. International Workshop on Requirements for Integrated Geographic Information Systems, New Orleans, Louisiana.

Austrian National Assembly (1975). Forstgesetz (Forestry law). **BGBl.Nr. 440/1975**.

Black, H. C. (1996). Black's Law Dictionary, West Publishing.

Brassel, K., F. Bucher, et al. (1995). Completeness. Elements of Spatial Data Quality. S. C. Guptill and J. L. Morrison. Oxford, Elsevier**:** 81-108.

Carver, S. (2001). Participation and Geographic Information: A Position Paper. Proceedings of the ESF-NSF Workshop on Access to Geographic Information and Participatory Approaches Using Geographic Information, Spoleto, Italy.

Federal Chancellery (2006). E-Government hat Priorität. Kurier. Vienna, Austria**:** 8.

Fisher, P. F. (1999). Models of Uncertainty in Spatial Data. Geographical Information Systems - Principles and technical Issues. P. A. Longley, M. F. Goodchild, D. J. Maguire and D. W. Rhind. New York, Wiley & Sons, Inc. **1:** 191-205.

Fisher, P. F. (2003). Data Quality and Uncertainty: Ships Passing in the Night! International Symposium on Spatial Data Quality, Hong Kong, Hong Kong University Press.

Frank, A. U. (2001). "Tiers of ontology and consistency constraints in geographic information systems." International Journal of Geographical Information Science **75(**5 (Special Issue on Ontology of Geographic Information)): 667-678.

Guptill, S. C. and J. L. Morrison, Eds. (1995). Elements of Spatial Data Quality, Elsevier Science, on behalf of the International Cartographic Association.

Guriev, S. (2004). "Red Tape and Corruption." Journal of Development Economics **73**: 489-504.

Hornsby, K. and M. J. Egenhofer (1997). Qualitative Representation of Change. Spatial Information Theory - A Theoretical Basis for GIS (International Conference COSIT'97). S. C. Hirtle and A. U. Frank. Berlin-Heidelberg, Springer-Verlag. **1329:** 15-33.

Medak, D. (2001). Lifestyles. Life and Motion of Socio-Economic Units. A. U. Frank, J. Raper and J.-P. Cheylan. London, Taylor & Francis. **8:** 139-153.

Navratil, G. (2004). How Laws affect Data Quality. International Symposium on Spatial Data Quality (ISSDQ), Bruck a.d. Leitha, Austria, Department of Geoinformation and Cartography.

Navratil, G. and A. U. Frank (2006, submitted). What Does Data Quality Mean? An Ontological Framework. AGIT, Salzburg, Austria, Wichmann Verlag.

Niemeier, W. (1985). Deformationsanalyse. Geodätische Netze in Landes- und Ingenieurvermessung II. H. Pelzer. Stuttgart, Konrad Wittwer**:** 559-623.

Peters, K. (1974). Problematik von Toleranzen bei Ingenieur- sowie Besitzgrenzvermessungen. Wien, Österreichischer Verein für Vermessungswesen und Photogrammetrie.

Reißmann, G. (1976). Die Ausgleichungsrechnung. Berlin, VEB Verlag für Bauwesen.

Salgé, F. (1995). Semantic Accuracy. Elements of Spatial Data Quality. S. C. Guptill and J. L. Morrison. Oxford, Elsevier**:** 139-151.

Searle, J. R. (1995). The Construction of Social Reality. New York, The Free Press.

Twaroch, C. (2005). Richter kennen keine Toleranz. Intern. Geodätische Woche, Obergurgl, Wichmann.

Veregin, H. (1999). Data Quality Parameters. Geographical Information Systems. P. A. Longley, M. F. Goodchild, D. J. Maguire and D. W. Rhind, John Wiley \& Sons, Inc. **1:** 177-189.

Wang, R. Y. and D. M. Strong (1996). "Beyond Accuracy: What Data Quality means to Data Consumer." Journal of Management Information Systems **12**: 5-34.

Wikipedia. (2006). "Leaning Tower of Pisa." From WikiPedia - The free Encylopedia Retrieved March 4th 2006, from http://en.wikipedia.org/wiki/Leaning_Tower_of_Pisa.

Zadeh, L. A. (1965). "Fuzzy Sets." Information and Control **8**: 338-353**.**

BIOGRAPHICAL NOTES

Prof. Andrew U. Frank

Prof. Frank is Professor of Geoinformation at the Vienna University of Technology since 1991. He teaches courses in spatial information systems, representation of geometric data, design of Geographic Information Systems for administration and business, selection of GIS software, and economic and administrative strategies for the introduction of GIS. In 1999 he became head of the newly founded Institute for Geoinformation and Land Surveying. In 2004 this Institute was merged with the Institute for Cartography and Geo-Media-Techniques and he is now the head of the Institute for Geoinformation and Cartography.

He leads an active research group focusing on problems of spatial cognition, user interfaces for GIS, and the economic and organizational aspects of the collection, management and use of geographic information. This work is supported by industry and research foundations. He is involved in several research projects of the European Commission. 1995 he completed a project for the organization of an international post-graduate course in GIS. Based on his experience from consulting assignments his present research interests comprise the cultural differences among European GIS-users as well as administrative and legal topics. He was also leader of the project “Study on European GI Policy Issues” under the GI2000 program of the European Commission. He is currently involved in EU IST projects.

Dr. Gerhard Navratil

Dr. Navratil is a research and teaching assistant at the Institute of Geoinformation and Cartography at the Vienna University of Technology. He graduated as a Surveying Engineer (Dipl.-Ing.) at the Vienna University of Technology in 1998 with a master thesis on the tasks of the Austrian Cadaster. In 2002 he received his Dr.techn. from the same University. His research interests are data quality, land administration and the problems combined with the historical development of land administration systems (geometrical as well as organizational problems).

CONTACTS

Prof. Andrew U. Frank
Vienna University of Technology
Institute for Geoinformation and Cartography
Gusshausstr. 27-29
A-1040 Vienna
AUSTRIA
Tel. +43 1 58801 12701
Fax. +43 1 58801 12799
Email: frank@geoinfo.tuwien.ac.at
Web site: www.geoinfo.tuwien.ac.at

Dr. Gerhard Navratil
Vienna University of Technology
Institute for Geoinformation and Cartography
Gusshausstr. 27-29
A-1040 Vienna
AUSTRIA
Tel. +43 1 58801 12712
Fax. +43 1 58801 12799
Email: navratil@geoinfo.tuwien.ac.at
Web site: www.geoinfo.tuwien.ac.at