Data Quality - Necessary Complement for GIS Based Decision Making *

H. Stanek
A.U. Frank
Department of Geoinformation
Technical University of Vienna
Gusshausstrasse 27 - 29
A - 1040 Vienna Austria
stanek@geoinfo.tuwien.ac.at
frank@geoinfo.tuwien.ac.at

ABSTRACT

The availability of data quality parameters improves decisions in earth resource information management. It is not sufficient that the necessary information is presented to the decision maker, but he must also have an appraisal of the quality of the information provided. This applies to all aspects of spatial decision making, from real estate ownership to environmental protection.
The first problem is to differentiate the basic components of data quality e.g. positional accuracy, attribute accuracy, temporal update level, scale etc. They are often compound in a lineage description, which is procedural (it provides information about the collection and processing of the data), but the goal must be an analytical description of data quality independent of processes used.
The other difficult problem is the assessment of the influence of data quality on the decision made. The paper reviews our current understanding of these problems based on specific case studies and lists items for future research.

INTRODUCTION

Quality of geographic information affects also decisions based on this data. Conventional processes can be investigated because of their quality requirements of the input data according to the planned work. Today available GIS should help to increase the economy effectively and minimize the time amount for spatial based planning processes. An other advantage is the combination of different data stored in GIS with analyses on combined data sources. So many application can be seen in general. Looking on details, a lot of criteria reduces the significant working area. One reason is the data capture process. Many data bases are still in the state of beginning because they are of enormous dimensions or the economic interest on it is very poor. Both reasons will coups incomplete data sets in present and obsolete data in future. For the further processing the description of data quality becomes more and more important.

The goal of this paper is to apply a given taxonomy of data quality parameters in special cases. If it fits in this cases a universal description of data quality this classification will also serve data quality in multiple used GIS applications. Cases which can be investigated on their analog conventional from of processing where chosen as case study field. Relevant procedures have to be analyzed and statements relevant for data quality can be extracted. The influence of specific properties like, long life span or highly correlation to a purpose, on data quality demands can be

* Presented at the 25th International Symposium, Remote Sensing and Global Environmental Change, Austria, 4-8 April 1993.

seen by aid of case studies. A comparison of unique data quality parameter from a different case studies view shows possible combining but also restrictions of multipurpose GIS. The consideration of data quality on GIS based decisions must obtain a guilty set of rules. The information about data quality can help the user in decisions about the further use of the product on several levels. First it is important for decision making to know about the quality of information it is based on. An other task is to generate quality information about the decision by the system itself.

CLASSIFICATION AND MEASUREMENT SCALES FOR DATA QUALITY PARAMETERS

That the structure of data quality in GIS can no longer be disregarded is demonstrated by the large number of publications on this specific topic. Its importance is also highlighted by the fact that the National Center for Geographic Information and Analysis has focused its Initiative 1 "Accuracy of Spatial Databases" on this issue. Goodchild (1992) offers a comprehensive and fundamental overview of existing attempts to structure and handle data quality parameters in databases. The work done in the NCGIA Initiative 1 concentrated primarily on the description and structuring of data quality and on its representation in databases.

In addition to modelling data quality in GIS - or as an integral part of a specific database - an efficient visualization of data quality is becoming an ever more important aspect. The NCGIA, realizing the need for further work in this field, started Initiative 7 "Visualization of Spatial Data Quality" (Beard, 1991), devoted to this topic. Also GIS education is now attaching more importance to data quality problems and the Core Curriculum of the NCGIA (Goodchild, 1990) also contains several units on "accuracy of spatial databases" and "managing error". In accordance with these issues in GIS the accuracy of databases can be structured as follows:

- positional accuracy
- attribute accuracy
- logical consistency
- completeness
- temporal update level

This structure defines a set of parameters for describing the taxonomy of data quality which are to the highest possible degree "independent", i.e. do not influence each other.

Today we have a number of standards which appropriately describe certain components of data quality. Parameters are also taken into consideration to an ever greater extent in the current efforts to define interfaces. Here the process of fine-tuning the five aforementioned parameters in such that they fit the purpose is a highly complex task. In particular if they are to be universally valid. The US. Spatial Data Transfer Standard, Data Quality Reporting Standard Specifications and British Ordnance Survey can be cited as examples. They all contain the general structure suggested by (Goodchild, 1992b). (Hunter, 1991) describes the typology of errors in spatial data bases as an interaction between causes, visualization and result error. From the point of view of the result these interactions are described in such a way as to make them universally valid. In his report about the situation (of standardization) in Australia, (Masters, 1991) specifically describes the work done on temporal update level and positional accuracy as well as their management for applications in operations specifically related to GIS.

It must be remarked that often a single parameter description for data quality would be preferred because of easy further processing. Sometimes the description in a single lineage parameter therefore was discussed (Frank, 1987). The problem is to find a definition based on a suitable value. As advantage in data acquisition, processing and also in decision making using this kind of managing data quality must be seen for a very specific application. The multipurpose usage of GIS

because of administrative and also economic reasons stands against this concept. It seems to be a good compromise to keep number of data quality parameters as small as possible.

The parameters can be distinguished from each other by looking at the modes of description applied to each of them. Positional accuracy and also statements about attribute accuracy are to be interpreted as a variance of an assumed distribution, which is often a normal distribution. Other statements, such as temporal update level, do not fit the mode of description of a function at all. The date of issue or of production is a fixed point in time or a period of time that defines this parameter in absolute terms.

Also our subjective assessment about quality statements may vary considerably. We feel more strongly about certain quality statements than about others. Comparisons with other magnitudes - such as the surface area of an apartment - might lead to wrong interpretations. This can be explained by the fact that people vary in their sensitivity to data quality or their need for data quality statements for linear or area-related magnitudes. .

CASE STUDIES

As the basis of such standardization efforts may serve the legal and technical provisions and regulations in force. These procedural regulations will have to be introduced into all existing systems where task processing has so far been carried out by conventional means and which are now switching to GIS. This conversion of existing task processing to GIS is a particularly interesting field. Obviously all the rules and regulations will have to be observed and complied with just as much as in the past. Beyond the mere replacement of conventional tools, the conversion to GIS is supposed to broaden the range of new applications. On the one hand, more complex analyses and modes of processing are made available to the user and, on the other hand, it will be possible to access, combine and share the data collection more easily for carrying out new tasks that were not envisaged before. However, the quality of this basic data soon demonstrates the formal limits of a GIS: it is of crucial importance to know about the quality of this data and also about the quality of the results achieved by working with this basic data collection. The case studies were selected because they appropriately relate to conventional modes of processing and their current transformation into GIS-applications. The choice is motivated in (Stanek & Frank, 1993).

In this paper we have attempted to assign the individual data quality parameters to various applications for which they are characteristic. We have relied upon the relevant technical and legal provisions or regulations in order to delineate the parameters from each other.

CADASTRE

The central unit of a cadastre systems or applications is the individual parcel or property. Governed by a series of rules and regulations, it has become an important vehicle for a complex and comprehensive data collection. The spatial references can be described by means of a geometrical figure with nodes and edges. The nodes represent boundary points that are marked by corner monuments in real nature and by means of coordinates in maps.

Many tasks based on cadastre data can be investigated with special regard to their data quality requirements. The division of a parcel located in land set aside for building is an important task based on the boundary cadastre. A few generally applicable regulations will be mentioned hereafter. Since building regulations in Austria are laid down by the provinces themselves, and not by the federal government, assignment parcel to an area or region is important. The spatial subdivision into cadastral communities helps to define this assignment.

In land set aside for building, parcels must have a certain minimum size. The shape of the divided-up building lots must be such that the newly created parcels can each be used separately for building purposes. Especially lateral distances between neighboring buildings or property boundaries frequently lead to controversies involving only a few centimetres. The space between structures is interpreted as a barrier that must not be infringed upon. Obviously, priority is here given to measurements that can be verified on site (surveying) over other values derived from computation. Therefore the values must be rather assigned to an attribute accuracy.

Whether the data should be assigned to attribute or to positional accuracy is often difficult to determine. This will be shown in another example. When a new building site is developed somewhere, the holders of the neighboring parcels will have to render certain services. One rule says, that they must cede part of their parcel up to the middle of the road for transfer into public property without financial compensation.

URBAN PLANNING

The focus here is no longer on the individual parcel but on the processing of several areas of investigation and on a comprehensive view of the entire planning region. Positional accuracy is no longer centered on the boundary point, but on entire areas or parts of the urban plan. Much emphasis is placed on how parcels are located in relation to each other and on the distance between two estates. In this context, positional accuracy is more aptly described as resolution. Resolution allows to visualize positional accuracy on the basis of the scale of the urban plan. This scale multiplied by an accuracy of the drawing of 0.1 mm defines a generalized statement on positional accuracy. All tasks in urban planning are based on a modelling of space often based on the results obtained by cartographic methods.

The methods of processing and reproducing maps or digitizing parts of maps have an effect on data quality. These cartographic methods need to be given consideration, besides analytical transformations or the choice of projections. The question is whether mathematical manipulations constitute ways of processing which can be repeated easily for part or the whole of the area that is represented. We can be sure that, besides their effects on positional accuracy, cartographic methods have an influence on the other data quality parameters as well.

Frequently, statistical data is tied to specific structures where used instead of original data. This is necessary, on the one hand, to reduce the data volumes, and, on the other, to allow inclusion of person-related data. The combined data are usable at least in an aggregated form by relating it to these structures like edifice rather than to individuals. This reliance on aggregated object data is necessary if a country's law forbid the direct use of person-related data.

NAVIGATION BASED ON NAUTICAL CHARTS

In this somewhat different field the study of data quality requirements reveals a number of particularly interesting aspects. The investigations here are based on the influences of cartographic processing techniques. Cartographic processing generally endeavors to achieve highly esthetically representations in order to make maps better readable and thus more user-friendly. In case of nautical charts other aspects are of domain interest. The chart depends strongly on its purpose.

Navigation based on nautical charts demonstrates further aspects of data quality: they record different navigational aids, such as beacon lights, dangerous depths, etc. These recordings should be absolutely accurate and complete, i.e. offer the highest degree of positional accuracy, on the one hand, and of attribute accuracy (depth, beacons, etc.) on the other. Here the required positional

accuracy is influenced by the value of the attribute. Thus, the parameter of completeness constitutes a key element for the definition of data quality in the navigation process.

Recency of the data is of primary importance. There is the requirement to have the charts updated by authorized agencies. The cause of an accident may have been a lacking update of the chart. The most relevant data quality parameters are completeness and temporal update level. A second aspect is planning the ship's course while taking into account positional accuracy (e.g. the ship has to navigate at a safe distance from the coast). In a large number of processes, recording the time and date of the most recent update is required by legal regulations (e.g. maximum admissible age of the map). In certain cases the date of a document may also point to the form of processing used when it was established or updated.
Different forms of nautical charts are used because of different demands on data quality parameters. Close to a harbor principal sketch gives much more relevant information than the general chart.

The use of a catalogue of laid-down symbols enhances the density of information to be achieved for the critical areas. Through color coding and restricted relations between different characteristics as part of the symbols the high quality requirements are met. Verification operations by means of procedures, that are as far as possible independent, is an absolute priority.

The following table summarizes characteristic items for the unique data quality parameter within the three case studies.

	Positional Accuracy	Attribute Accuracy	Logical Consistency	Completeness	Temporal Update Level
Cadastre:	• Positional accuracy of a single boundary point • Agreement of neighbors on the boundary line	• Ownership, correct by law • Land use • Address (not updated)	• Enforced by verification • Land register authority • Supplementary legal provision	• Verification procedure • Complete coverage • Complete description of ownership • Incomplete additional information	• Update indicator • Daily update – temporal granularity • Date of inquiry • Title deed collection
Urban Planning:	• Accuracy of drawing • Resolution • Generalization • Partially varying accuracies • Varying methods for data capture and processing	• Methodically influenced by aggregation of data	• Preliminary testing in small areas by taking random samples	• Statistical techniques • Independent verification	• Heterogeneous temporal references on account of different forms and their data capture
Navigation:	• Varying requirements within the chart • Verification by means of independent determination • Error avoidance due to easy-to-use strategies	• Location-related requirements – only task relevant • Redundancy	• Obtained through standardized modes of description • Assured by color codes etc. • Symbols: representing absolutely complete data sets	• High or absolute requirements only for critical areas • Highly time-sensitive for specific items	• Agency authorized to carry out updates • Changes verifiable • Future changes as forecast

DECISION MAKING AND DATA QUALITY MANAGEMENT

The possibility to utilize or share data within a GIS for several different applications is often one of the main reasons for acquiring a GIS. However, such data sharing and multiple data use is in actual fact often restricted by law and also by costs. A combination of the cadastre case study and the urban planning study may serve as an example.

In particular in densely populated areas, the tasks of regional planning often merge with those of local urban planning. Such a system that tries to merge these fields is being developed for the municipal administration of the city of Vienna (Belada, 1990). Based on a digital multipurpose map at a basic scale of 1:200 a data base and map-system is being evolved to eventually result in a combined information system for planning and administration.

For a planned division of land set aside for building it is for instance required to respect the provisions laid down in the building regulations. The information about building regulations that relates to one particular parcel are obtained from a regional urban plan. The close interweaving of the legal aspects already requires a detailed analysis of data quality parameters. This is why the conversion to a GIS-based system requires the definition of unequivocal rules for task processing.

Other regulations define the adjacent properties in terms of the influences acting upon them because they are located e.g. in the main wind direction, as well as taking into account the distance between estate and site. The position of the estate must be determined beforehand by aggregation and then verified as to its distance to the site. Since all the owners of the adjacent properties have to be invited to participate in the building negotiations, their number can sometimes total several hundred persons. If only one of them was left out and not invited, this can result in a new verification of the building permit granted. This may result in an injunction ordering - depending on the construction stage - a temporary stop of construction work or even a temporary shutdown. This again demonstrates the high sensitivity with regard to completeness and positional accuracy.

The case study of nautical charts has a distinctly different structure. There is a close relation to the purpose of the system. Because of the spatial dependence of the individual data quality parameters, a transfer in a GIS seems at first sight unreasonable. However, if we aim at the purpose of the navigation process in its entirety, the situation becomes a different one. The purpose lies in the quick and simple determination of a geometrical location with the assistance of landmarks and the determination of the course whilst taking into account danger zones, nautical rules, regulations and guidelines as well as economic aspects. Nautical charts are designed to measure and process angular observations obtained in taking one's bearings. The selected map projection is often a representation in which the angles are absolutely true to nature, in order to avoid reductions of measurements. Working out the position is an inversion of the observation procedure.

Determining the position on the basis of longitude and latitude for the above described purpose is not immediately necessary. When GPS receivers and other externally based procedures are used for navigation, the position must be plotted on the map in longitude and latitude. Taking over these methods into a digital navigation system avoids errors in plotting the position and enhances the security of the entire navigation process. This system is also able to check the keeping of safety distances or minimum water depths. In this process also statements about data quality, such as positional accuracy and attribute accuracy will have to be considered. Besides the position statement, also the nautical chart and additional information must be available in digital form. Since responsible navigation must always allow an independent control, beacons, landmarks etc. will also have to be included in the navigation process also in the future. Systems will have to be able to process simple forms of observation such as goniometric plotting also in the future. For navigational tools there is a high requirement for recency of information.

SUMMARY

The investigations of data quality in GIS that have been carried out so far were primarily aimed at the description and analysis of accuracies. The main object of analysis has been positional accuracy and, in specific cases also attribute accuracy. Which other parameters are to be taken into account, i.e. consistency, completeness, or lineage will depend on the envisaged application. The proposed classification of data quality into five parameters is appropriate to meet the requirement of a complete description of data quality in the fields investigated by case studies. The taxonomy of data quality proposed in GIS in five independent parameters can be verified by analyses of operational instructions of conventional processing.
The analysis of conventional processing sequences, in which also their data quality parameters are included, should improve modelling. Special GIS applications can, as far as their requirements of data quality are concerned, be grouped together by means of a scale of reference. For a combined application of the tasks of case study cadastre and urban planning, the combined requirements of data quality can be derived as a combination of the contents

The use of one data collection for different tasks requires an exact and comprehensive analysis of the legal rules and regulations that are applicable for the processing of each of the particular tasks. For many areas, in which GIS is to be used in the future, the provisions and regulations to be observed already exist today. They can be transformed such as to enable the user to define the differing requirements of data quality to be met for the specific process he works on. This can at least be used as a working basis for the description of overlapping applications like managing building regulation or similar tasks.

Management of data quality is not a new thing coming up with GIS. Data Quality requirements must be regarded from the expected field of application. A set of rules and regulation says more or less details about the data quality. So the decisions made based on this data are confident and verifiable. The conversion of existing decision making procedures to GIS causes the investigation of data quality from the source over data processing methods to the final decision making process. This comprehensive data quality management has to be introduced in GIS for confident decision making.

REFERENCES

Beard, K., M., Buttenfield, Barbara and Clapham Sarah B. (1991). NCGIA Research Initiative 7: Visualization of Spatial Data Quality Scientific Report for the Specialist Meeting 8-12 June 1991, No. Technical Paper 91-26.). Castine, Maine" National Center for Geographic Information and Analysis.

Belada, P. (1990). Die "Mehrzweckskarte" der Stadt Wien. Österreichische Zeitschrift für Vermessungswesen und Photogrammetrie, 78(Heft 3 / 1990), 106-123.

Frank, A. (1987). Overlay Processing in Spatial Information Systems. In AUTO CARTO 8, Eigth International Symposium on Computer-Asssisted Carttography, Baltimore, MD, ASPRS & ACSM.

Goodchild, M. F., Kemp, Karen, K.(ed.) (1990). Technical Issues in GIS - NCGIA Core Curriculum No. units 45, 46). National Center for Geographic Information and Analysis.

Goodchild, M. F. (1992a). Accuracy of Spatial Databases (Final Report initiative 1) National Center for Geographic Information and Analysis.

Hunter, G. J. (1991). Processing Error in Spatial Databases: The Unknowen Quantity. In H. G. J. (Ed.), Symposium on Spatial Database Accuracy, (pp. 203 - 214). Melbourne, Australia June 19-20, 1991: Department of Surveying and Land Information The University of Melbourne.

Masters, E. (1991). Defining Spatial Accuracy. In G. G. Hunter (Ed.), Symposium on Spatial Database Accuracy, (pp. 215-224). Melbourn, Australia:

Stanek, H., Frank, A. (1993). GIS Based Decision Making Must Consider Data Quality. In EGIS, . Genoa, Italy March 29 - April 1 : (pp. 685-692)

GRAZ
Our way of being a town.

25th International Symposium

Remote Sensing and Global Environmental Change
Tools for Sustainable Development
4-8 April 1993

PROCEEDINGS

Volume II
Interactive Poster Sessions

CIESIN
Consortium for International Earth Science Information Network

JOANNEUM RESEARCH

ERIM