To Characterise and Measure Completeness of Spatial Data: A Discussion Based on the Digital Chart of the World (DCW)

Sindre Langaas

UNEP/GRID-Arendal,
c/o Dept. of Systems Ecology, Stockholm University, S-106 91 Stockholm, Sweden.
Fax: +46-8-158417 Phone: +46-8-161737 E-mail: langaas@grida.no

Håvard Tveite

Department of Surveying,
Agricultural University of Norway, P.O.Box 5034, N-1432 Ås, Norway.
Fax: +47-64948856 Phone: +47-64948857 E-mail: lanht@nlh.no

Abstract

There is an increasing degree of sophistication associated with describing the qualities of spatial data. Completeness is one data quality component that is included in both the US Spatial Data Transfer Standard (SDTS) and the European standard under development within the framework of the European Committee for Standardisation (CEN). While both standards use the same term, there are apparent semantic differences reflected in their definitions and proposed ways for assessment and reporting. This paper will discuss these differences and their implications using the global 1:1 million scale Digital Chart of the World (DCW) database as a test case.

Keywords: Completeness, SDTS, DCW, spatial data quality

1. Spatial data quality characterisation and measures

The change from paper maps to GIS **data** in various kinds of geographical data analysis and applications has made it easy to use the same spatial data for different applications and also for combing several layers into quite complex spatial models. This has created a need for data quality descriptions and measures to be attached to the datasets (whatever definition used of dataset). Thereby the user can judge the suitability for an intended application. A commonly used definition of quality is ‘fitness for use’ (Chrisman 1984). Further, if several spatial datasets with appropriate quality measures are combined, the error propagation can be modelled. Here Veregin’s hierarchical error model comes to mind (Veregin 1989).

In Figure 1 is shown a modified version of Veregin’s (1989) well-known model of the hierarchy of needs in modelling of error in GIS operations. We have adapted this model to more recent terminology and found that a hierarchy of needs for handling of spatial data quality better reflects the current status in terminology. In this model, level 1 is concerned with classification and identification of spatial data qualities. The efforts dedicated to the classification of spatial data qualities are reflected in the data quality parts of several on-going standardisation efforts. Level 2 focus on the characterisation and assessment of the qualities defined in level 1.

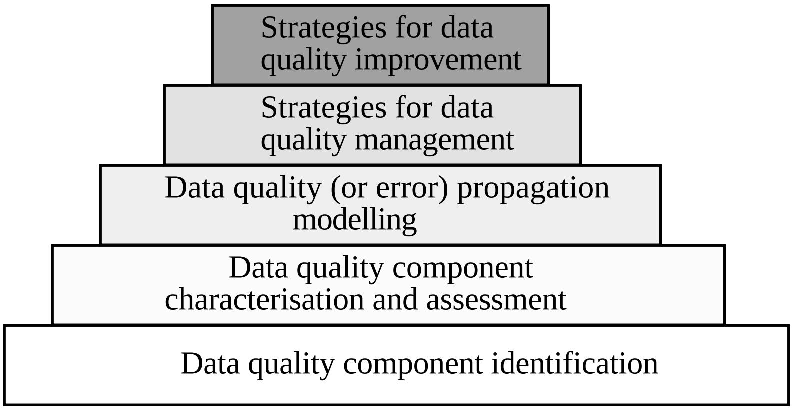

Figure 1. A hierarchy of needs for handling of spatial data quality. The text and concepts in this figure, based upon Veregin (1989), is modified according to recent terminology

Several countries or groups of countries within both the civilian and military sector of the spatial data community have for a number of years worked on standards to facilitate transfer and use of spatial data (Moellering 1992). Two major efforts are the US Spatial Data Transfer Standard (SDTS, Fegeas *et al.* 1992) and the European standard currently under the development of European Committee for Standardisation (CEN) Technical Committee 287 (CEN/TC287/WG02 1995). These standards include data quality components.

SDTS including Part 1 with the data quality report specifications was approved in July 1992 and is currently being implemented by federal, state and private spatial data producers in the USA. The European Geographic Information standard with its Data Description - Quality part is currently being developed and is supposed to be ready by 1997/98. Both standards have, within the data quality part of their specifications singled out a quality component termed *completeness***.** While both standards use the same term, there are apparent semantic differences reflected in their definitions and proposed methods for characterisation and assessments.

In this essay we will briefly describe and discuss some of these differences. We will do so in view of experiences from an on-going project aimed at reporting of data quality of the Digital Chart of the World (DCW, ESRI 1992, Langaas and Tveite 1994). We want to highlight some aspects relevant to the usefulness of the two different completeness concepts and their suggested reporting characteristics and measures.

2. Completeness - reporting characteristics, measures and metrics

2.1 SDTS

The term *completeness* is not defined explicitly in the SDTS. It is stated, though, under the completeness section that 'the quality report shall include information about selection criteria, definitions used and other relevant mapping rules.' Further, 'The report shall describe the relationship between the objects represented and the abstract universe of all such objects. In particular, the report shall describe the exhaustiveness of a set of features. Exhaustiveness concerns spatial and taxonomic (attribute) properties, both of which can be tested.' The concept 'abstract universe of all such objects' is a key concept which in each case needs an accurate definition (or specification) to give the necessary information about the various completeness aspects.

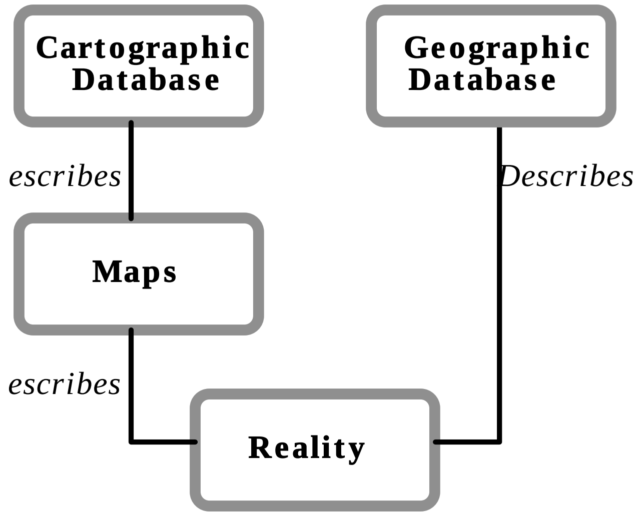

Figure 2. While geographical databases model and describe reality directly, cartographical databases do this indirectly.

In these specifications of completeness characteristics it appears that a cartographical digital database (or dataset) rather than a geographical digital database has been in mind. The distinction between cartographical and geographical databases is visualised in Figure 2. Here it is seen that ‘reality’ for cartographical databases are modelled twice. First, to create maps using not only strict objective thematic criteria but also cartographical criteria for readability and aesthetic purposes, and secondly these map(s) are modelled to derive a digital database. In a conventional thematical map production process there exist a wide range of selection criteria, specific definitions and other mapping rules that convey information about the suitability of the digitised version also for other purposes than the initial thematic one.

In the SDTS, completeness reporting is primarily supposed to be done as textual reports and to a lesser extent as quantitative measurements, although it is referred to objective tests that can be carried out.

2.2 CEN/TC287/WG02

In CEN/TC287/WG02 (1995) completeness is defined as 'the difference between an actual dataset and its specifications.' It is further stated that 'completeness measures indicate how well the information reflects the content defined by the specification'. Taking this quantitative approach, three possible measures are suggested to quantify completeness. These are *omission, commission* and *coverage ratio***,** represented by the following metrics:

- Percentage of data missing relative to specification,
- Percentage of data present that is not in current specification of dataset or extract, and
- Occurrences of one variable per unit of another.

The CEN/TC287/WG02 assessment approach is more concise compared to SDTS. However, given its definition and recommended approaches of assessments, being solely quantitative, the precise definition of 'dataset' (what is a dataset ?) and 'specifications' given in quantitative terms are crucial for implementation. Furthermore, while the recommended approaches appear well suited for geographical datasets (or databases), they are more difficult to implement for cartographical datasets.

3. Completeness reporting of the DCW - some considerations

3.1 DCW - a cartographical database

The completeness quality aspect is of particular relevance for DCW. The digital DCW was made from two map series, the Operational Navigation Charts (ONC, 1:1 mill.) and the Jet Navigational Charts (JNC, 1:2 mill. Antarctica only). The DCW is a digital representation of the ONCs and JNCs and therefore a cartographical database. Both map series are made with a large number of mapping rules reflecting their intended purposes (DMA 1981):

"The 1:1,000,000 scale Operational Navigation Charts (ONC) Program provides aeronautical charts to support medium altitude enroute navigation by dead reckoning visual pilotage, celestial, radar, and other electronic techniques. In the absence of Tactical Pilot Charts (TPC's), these charts should also satisfy the enroute visual/radar navigation requirements of pilots/navigators flying low altitude operations (500 feet to 2000 feet above ground level). The ONC is also used for operational planning, intelligence briefings, and preparation of visual cock-pit displays/ film strips essential to aerospace navigation of high-performance weapon systems."

3.2 Completeness - the issue of 'ideal' reporting level exemplified

The quality reporting ideally should be assigned to various levels of the dataset. SDTS distinguishes between the following levels:

- Dataset (or database)
- Theme
- Map (or geographical extract)
- Feature/object (or thematical extract)
- Element

To clarify the difference between these levels, an example will be given.

An environmental researcher would like to use the DCW to quantify potential annual increase in methane (CH_4) releases from *cranberry bogs* in Northern Finland under doubled atmospheric CO_2 levels and associated temperature rise. *Cranberry bog* is one class or feature under the layer (or theme) Land Cover in the DCW. Completeness descriptions on the entire <u>dataset</u> level might be of limited relevance. However, the knowledge on the specific purposes of the ONC and JNC map, (i) aerial navigation and (ii) military strategic planning, obviously indicates that the information contained on cranberry bogs might be unsatisfactory. The next level of reporting is the <u>theme</u> level. Cranberry bog constitutes one class or features out of many in the theme (or layer) Land Cover of the DCW. A completeness description on the theme level, supposedly valid for the spatial extent of the entire datasets, then will provide more detailed information about the suitability for annual methane emissions. The next level of reporting might be the feature/object level. If specific completeness information is available on the cranberry bogs *per se*, then the researcher would be even better prepared to evaluate the suitability of the DCW for its planned application. Although not so relevant in this case, one might also find that completeness reporting down to the element level can be provided. Depending upon the spatial coverage of the dataset in question, the completeness reporting on the four levels; (i) dataset, (ii) theme, (iii) feature/object and (iv) element ideally should be provided for specific regions. Individual map sheets in the ONC or JNC charts are an obvious sub-division of the entire dataset region into smaller specific regions for reporting. Evidently, completeness reporting on cranberry bogs on the feature/object level for those map sheets that cover Northern Finland would be the most specific and useful completeness reporting that could be provided.

3.3 Quantitative completeness assessments of DCW

The SDTS recommends a topological test as the only quantitative (or structured) approach for completeness assessments besides textual reporting. In the European standardisation efforts, the quantitative approaches are the only ones recommended. Could these be employed for the DCW ? At the database level - hardly. 1.5 GB of digital data effectively prohibits this. At the theme level - hardly, but more feasible if reference data is available. When coming down to the geographical and thematical extract levels this becomes, theoretically at least, more attractive. Completeness assessment using the suggested measures *omission*, *commission* and *coverage ratio* requires (i) precise and quantitative specifications and (ii) relevant reference data that are presumed to be of a higher quality. Higher quality in this context means that the reference data comply better with the specifications given for the various themes and geographical regions. Such reference datasets are virtually non-existent given the purpose of the original map series referred to earlier and the associated detailed mapping specifications described in DMA (1981). One can, however, apply other existing and more general purpose geographical datasets that thematically are quite similar to the themes of the DCW. From a user perspective different from the initial ONC and JNC purpose - mirrored in the DCW database - this is quite satisfactory. Most users of the DCW are not using it for the aerial navigational and military planning purpose. Therefore, for the example given in para. 3.3 quantitative assessment with the recommended measures omission and commission for cranberry bogs in Northern Finland, provided that such digital data of high quality exist are feasible, would be highly attractive to the environmental researcher. However, this assessment would not give 'the difference between an actual dataset and its specifications.' The specifications as given in DMA (1981) is:

> “Rice fields, cranberry bogs and “similar flooded areas” shall only be shown when they are very unique or distinctive features in areas devoid of landmark detail.”

It is obvious that this specification is very subjective and renders testing almost impossible .

3.4 DCW completeness reporting - what do we do ?

Within our DCW Data Quality project we have chosen the proposed SDTS completeness understanding and approach for assessment reporting. This is more directed towards cartographical databases than the completeness part of the data quality section of the European standard under development. In practice, this means to summarise and structure the definitons and specifications given in DMA (1981). It should be mentioned though that Lineage and Usage part of the European standard does allow for extensive textual information. The completeness information or actual cartographical mapping rules instead can be reported in these parts.

Acknowledgements

This work has been partially funded by the Norwegian Research Council under the Geographical Information Technology Programme (Project No. 31783/ 223) and UNEP/GRID-Arendal (Project No. 205026)

References

Aalders, H. J.G.L., A. Giordano, and O. Jacobi. 1995. Quality: Discussion paper for an ESF GISDATA conference on quality. Published by the GISDATA Secretariat, Dept. of Town & Regional Planning, University of Sheffield, UK.

CEN/TC287/WG02. 1995. Geographic Information - Data Description - Quality. Draft for discussion produced by project team 5 (PT05), 1995-1-24, CEN Central Secretariat, Brussels, 36 pages.

Chrisman, N. 1984. The role of quality information in the long-term functioning of a Geographic Information System. In: Proceeding of AUTOCARTO 6, Vol. 2, ASPRS, Falls Church 1983, pp. 303-321.

DMA. 1981 - . Product specifications for Operational Navigation Charts (Code: ONC) Scale 1:1,000,000. First edition 1981 and changes and amendments thereto. Defence Mapping Agency, Washington., D.C.

ESRI. 1992. The Digital Chart of the World for use with ARC/INFO® Data Dictionary. ESRI, Redlands, CA.

Fegeas, R.G., J.L. Cascio, and R.A. Lazar. 1992. An overview of FIPS 173, The Spatial Data Transfer Standard. *Cartography and Geographic Information Systems* **19**(5): ??-??.

Langaas, S. and H. Tveite. 1994. Project proposal: Issues of error, quality and integrity of digital geographic data: The case of the Digital Chart of the World. URL file://ilm425.nlh.no/pub/gis/dcw/quality.ps

Moellering, H. 1992. *Spatial data base transfer standards: Current international status.* International Cartographic Association Commission on data standards, Elsevier Applied Sciences, NY.

Tveite, H. and S. Langaas. 1995. Accuracy assessments of geographical line datasets, the case of the Digital Chart of the World. In this Volume.

Veregin, H. 1989. Chapter 1. *Error modelling for the map overlay operation.* In: M. Goodchild and S. Gopal (eds.), Accuracy of spatial databases, Taylor & Francis, pages 3-18.