Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use Within GIS

Rodolphe Devillers, Yvan Bédard, and Robert Jeansoulin

Abstract

Metadata should help users to assess the quality (fitness for use) of geospatial data, thus reducing the risk of data misuse. However, metadata presents limitations and remain largely unused. There still exists a need to provide information to users about data quality in a more meaningful way. This research aims to dynamically communicate quality information to the users in a rapid and intuitive way in order to reduce user meta-uncertainty related to geospatial data quality, and then reduce the risks of data misuses. Such a solution requires a data model able to support heterogeneous data quality information at different levels of analysis. Using a multidimensional database approach, this paper proposes a conceptual framework named the Quality Information Management Model (QIMM) relying on quality dimensions and measures. This allows a user to easily and rapidly navigate into the quality information using a Spatial On-Line Analytical Processing (SOLAP) client-tied to its GIS application. QIMM potential is illustrated by examples, and then a prototype and ways to communicate data quality to users are explored.

Introduction

The context in which geospatial data is used has changed significantly during the past decade. Users now have easier access to geospatial data and GIS applications, especially through the internet. As they were formerly almost restricted to geospatial experts, it is now frequent that users with a limited expertise in the geospatial domain use geospatial data and GIS applications. Although this is a positive evolution in general, one problem has emerged: today's typical geospatial data users have less knowledge in the geographical information domain (Agumya and Hunter, 1997; Aalders and Morrison, 1998; Curry, 1998). Consequently, their knowledge about risks related to the use of geospatial data is limited (Goodchild, 1995; Agumya and Hunter, 1997; Curry, 1998; Elshaw Thrall and Thrall, 1999). In that sense, Goodchild (1995) argues that "GIS is its own worst enemy: by inviting people to find new uses for data, it also invites them to be irresponsible in their use". This sometimes leads to faulty decisions based on these data, possibly having significant social, political or economical consequences which is being discussed in the literature (Beard, 1989; Monmonier, 1994; Curry, 1998; Agumya and Hunter, 2002; Gervais, 2004). In order to reduce the risks of misuse, geospatial data producers spend a lot of resources documenting their datasets to inform the users on dataset specifications and quality. Among these documents, metadata (i.e., data about data) provide information on several aspects of the datasets, such as, data producer identification, spatial reference systems, lineage, definition of features or attributes, and data quality (FGDC, 2000; ISO-TC/211, 2003). However, metadata are defined in the literature as producer-oriented offering only limited benefits for the users who want to assess the fitness of the data for their use (Frank, 1998; Harvey, 1998). In fact, experience shows that metadata do not reach their information goal for non-expert users and are also difficult to understand by many expert users (Timpf *et al.*, 1996; Frank, 1998; Harvey, 1998). Understanding and reaching conclusions that could be used legally, for example, about the quality of geospatial data rapidly becomes an unmanageable task when one wants to take into consideration the spatial, temporal, thematic, acquisition, and other heterogeneities found in a dataset. Consequently, metadata related to data quality usually remain unused by non-expert as well as by experts, even with the best datasets, leaving users in a state of ignorance about the characteristics of the geospatial dataset being used.

As demonstrated by Gervais (2004), an increasing number of geospatial data is intended for general public and must follow legal requirements related to a mass-product category. Metadata, as currently provided or defined within international and national standards, do not reach these obligations, especially concerning the requirements of providing easily understandable information as well as information about potential risks of misuse. According to Gervais (2004), there is a need for a computerized instruction manual that would reduce the risks of misuse by providing to the users of geospatial data information that is easier to understand. Several authors highlighted the need to design such a tool, sometimes identified as "Quality-aware GIS," "Quality GIS," or "Error-aware GIS" that would dynamically take quality information into consideration during data manipulation (visualization, queries, and updates) in order to prevent the user from "illogical operations" (Unwin, 1995; Hunter and Reinke, 2000; Duckham and McCreadie, 2002; Qiu and Hunter, 2002).

R. Devillers and Y. Bédard are with the Centre de Recherche en Géomatique (CRG), Université Laval, Québec, G1K 7P4, Canada, (rodolphe.devillers.1@ulaval.ca; yvan.bedard@scg.ulaval.ca).

R. Jeansoulin is with the Laboratoire des Sciences de l'Information et des Systèmes (LSIS), CMI, Université de Provence, 39 Rue Joliot Curie, 13453 Marseille Cedex 13, France (robert.jeansoulin@cmi.univ-mrs.fr).

Photogrammetric Engineering & Remote Sensing
Vol. 71, No. 2, February 2005, pp. 205–215.

0099-1112/05/7102–0205/$3.00/0

Such systems require automatical accessing and use of the information related to geospatial data, i.e., metadata. Such metadata do not have to be restricted to the metadata identified or provided by different standard organizations or data producers, they can refer to "data about data" in a more general way. However, today's systems have not yet achieved an efficient user-centric management of geospatial data quality information. The goal of this paper is to propose a conceptual framework for the management of geospatial data quality information that aims to go one step ahead of existing solutions.

In the next section, we explain how this research fits into the wider evolution of geospatial data transfer especially focusing on today's practice of making metadata accessible to users for assessing the fitness for use of their datasets. Then, we present the state of the art concerning "what" kind of quality information is available today. We do so by presenting different standards and classifications of data quality information. The following sections present different hierarchies allowing quality analysis at different levels of detail and based on the literature, we propose a conceptual framework for geospatial quality information management. We describe multidimensional data structures as well as Spatial On-Line Analytical Processing (SOLAP) and discuss their relevance for geospatial quality information management. A framework for a SOLAP model managing data quality information is presented. We then illustrate our approach with different scenarios of user navigation within the quality information model. We finally present our prototype based on the quality information model developed to test the concepts, and highlight the impact of such a model on quality information communication.

Issues About Geospatial Data Transfer and Quality

In the past, geospatial data were typically produced and used within the same organization. Knowledge about data production processes and characteristics, including quality, was more implicit (i.e., organizational memory) than explicit (e.g., metadata). With the arrival of digital data, the increase of data transfer changed this perspective. The way organizations or people communicate information related to geospatial data followed an evolution during which information transferred became more accessible or meaningful to a larger group of geospatial data users (Figure 1). Several stages can be identified:

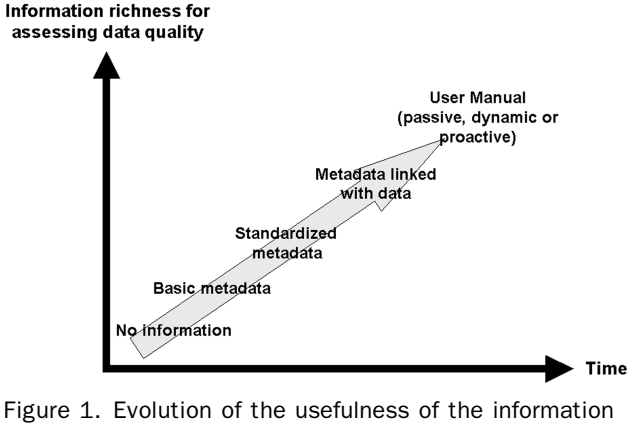


Figure 1. Evolution of the usefulness of the information communicated to data users for assessing geospatial data quality.

(1) No quality information: Data is distributed without any associated metadata describing them. This situation is still very frequent, and it is not rare to see users specifically asking for the metadata not to be sent, even when they do exist;

(2) Basic quality information: Data producers provide some information when distributing their geospatial datasets, such as dataset reference systems, spatial accuracy, or production date. However, this information, not compliant to any standard, is different from one organization to another describing different characteristics of the datasets at various levels of richness;

(3) Normalized metadata: Local, national, or international organizations, such as ISO/TC 211, OPENGIS, FGDC, CGSB/COG or CEN/TC 287, suggest geospatial metadata standards in order to homogenize the information shared between the organizations. However, different standards can be used from one organization to another. Often stored in separate text files, these metadata are rarely explicitly associated with their related data, limiting their usefulness for GIS functions (e.g., associating uncertainty to distance measurements based on positional accuracy metadata). Furthermore, these standards are more producer-oriented than user-oriented, being more a formalization of production procedures and tests understandable by data acquisition specialists, than meaningful information for a general audience useful for decision-making processes;

(4) Metadata linked with data: Although metadata provided with datasets are still regularly stored in a text file separate from their data file, without any explicit link between both, some research projects both in academia and industry are now being performed to strengthen the link between metadata and the data they describe, up to the instance and attribute levels. Beard (1997) mentions "there is potentially great benefit from an integral association of data with descriptions or measures of its quality. Approaches which separate quality descriptions from the data risk reducing ease of access." Such structured quality information would be easier to access by users or software programs, but more difficult to generalize if the granularity of quality information is very fine. One of the reasons for a tighter link is the need to propagate data updates to metadata. An explicit link between metadata and data would also allow the dynamic use of metadata during data manipulation. Commercial tools, such as, ArcGIS® ArcCatalog® (ESRI), or SMMS® for Geomedia® (Intergraph), provide a way to manage metadata and dynamically link them to data. However, these tools are still limited in terms of the types of metadata that can be stored and the level of detail of the metadata (i.e., metadata are usually stored on the dataset or object class level only).

We suggest a stage further exploiting the metadata structured in the Stage 4. This level, exemplified by the Multidimensional User Manual (MUM) project (Devillers *et al.*, 2002), provides high-level information or functionalities aiming at reducing the risks of misuse by reducing users' meta-uncertainty when manipulating geospatial data.

(5) The User Manual can be divided into three complementary parts, namely Passive, Dynamic and Proactive User Manual.

- Passive User Manual: The Passive User Manual is defined as a textual User Manual as usually provided with other goods (e.g., medical drugs, electronics), providing different information related to dataset specifications, possible use, and limitations. Such a manual can rely on metadata, other information, or recommendations provided by data producers or experience from other parties that had used these datasets in a certain context. Each manual is contextual, being produced for certain data being used in a certain context.
- Dynamic User Manual: The Dynamic User Manual is designed to be integrated within a GIS interface. Such a manual provides users with relevant aggregated information and allows them to navigate at different levels of detail within this information (Devillers *et al.*, 2002). Using different levels of detail helps to avoid informa-

tion overload and to synthesize the quality information. The information provided to the user is either quantitative or qualitative (the latter being more frequent at general levels, while the former more frequent at detailed levels) and would help identify some datasets characteristics that could possibly be risky for the intended use. Doing so requires comparing users' expectations with intrinsic characteristics of geospatial data.

- Proactive User Manual: The Proactive User Manual is designed to act directly on-the-fly on users' GIS operations in order to avoid some data misuse. This stage requires a database of "illegal operations" as described by Hunter and Reinke (2000). Based on this knowledge and the metadata, the system could also avoid the use of certain functions in some contexts or display a message to warn the user on the possible consequences of the action (e.g., restrict data visualization to certain scales based on the data acquisition scale; associate uncertainty to calculations results: e.g., distance measurement).

This paper focuses on Stage 4 presented above, which describes how to link metadata and their associated data to allow the User Manual, or any other "Quality-aware GIS" functions, to work properly. This work constitutes the basis on which the Stage 5 relies. For the scope of this paper, quality information is defined as any information allowing the assessment of dataset quality (fitness for use). Then, quality information includes metadata provided with datasets but may also include other relevant information or even expert opinions about given data.

Geospatial Data Quality Characteristics

The definition of a data model allowing the management of geospatial data quality information requires knowing *what* quality information is available and can be integrated into such model. This section provides an overview of the literature in terms of data quality classifications looking at both metadata standards and academic research in order to highlight the diversity and similarities of quality classifications, present the limitations of metadata, and then support the QIMM model to be described.

Data quality issues have been extensively explored in the geographic information domain for about 20 years. However, definitions on the meaning of "quality" remain various. Two trends can be identified in the literature. One restricts quality to dataset internal characteristics, i.e., intrinsic properties resulting from data production methods (e.g., data acquisition technologies, data model, and storage). This trend is often identified as internal quality. The other trend follows the "fitness for use" definition (Juran *et al.*, 1974; Chrisman, 1983; Veregin, 1999), quality being defined as the level of fitness between data characteristics and user's needs. This trend is often identified as external quality. As opposed to the former trend, the latter sees quality as a concept that is relative to the users and usages, neither an independent nor absolute concept. External quality assessment requires information describing internal quality, the concept of external quality then being larger than the internal one. Works suggesting classification of geospatial data quality information are typically approached from two different perspectives: producer and user. Producer point of view generally focuses on internal quality, while user point of view looks at both internal and external quality.

Several quality characteristics are suggested by standards organizations and academic researchers for defining both internal and external qualities. Standardization bodies largely developed the data producer perspective (e.g., CEN/TC 287, ICA, ISO/TC 211, OPENGIS, SDTS). They usually classify data quality into five to seven parameters being: lineage, positional accuracy, attribute accuracy, semantic accuracy, temporal accuracy, logical consistency, and completeness (CEN/TC-287, 1994/1995; Guptill and Morrison, 1995; FGDC, 2000; ISO-TC/211, 2003). Each class is usually composed of several sub-classes, but few of these address issues such as accessibility (costs, delays), rights to reproduce (copyright policy), official or legal character of the data, privacy restriction, or any other issues that are needed to assess the fitness for use (from the user's point of view). Table 1 provides an overview of geospatial data quality characteristics identified in standards (i.e., CEN, ICA, ISO and SDTS) or by a data producer organization (i.e., IGN-France). This table reflects the meaning of quality characteristics (i.e., if two organizations have two different names for similar aspects of the quality, they are grouped in a same category).

TABLE 1. EXAMPLES OF DATA QUALITY CHARACTERISTICS PROVIDED BY STANDARDS OR CARTOGRAPHIC ORGANIZATIONS

	CEN[1]	ICA[2]	IGN[3]	ISO[4]	SDTS[5]
Lineage/Source	X	X		X	X
Spatial/Positional Accuracy	X	X	X	X	X
Attribute Accuracy				X	X
Semantic Accuracy	X	X	X	X	
Completeness	X	X	X	X	X
Logical Consistency	X	X	X	X	X
Temporal Information/ Accuracy	X	X		X	

[1](CEN/TC-287, 1994 and 1995), [2](Guptill and Morrison, 1995), [3](IGN, 1997), [4](ISO-TC/211, 2003), [5](FGDC, 2000)

Table 1 shows that standards and data producers (1) mainly focus on *internal quality* (e.g., accuracy, completeness, and consistency) aspects, and (2) agree, in general, on similar characteristics. Standards are now generally converging to the ISO international standard that may serve as reference for the quality characteristics identification.

On the other hand, different authors argue that quality assessment defined as "fitness for use" may require information that is not yet included in geospatial metadata standards. They suggest quality characteristics in the wider approach of external quality (i.e., quality in the context of use) in addition to internal quality. For instance, Aalders and Morrison (1998) add to the ISO criteria information related to data usage being previous use of a dataset by other users for various applications (i.e., organization that has used the dataset, type of usage and its perceived fitness, possible constraints, or limitations during the use). Bédard and Vallière (1995) bring other characteristics such as legitimacy (legal or *de facto*) and accessibility (costs, delays, easiness to obtain) of the data. Working on data quality issues in general (i.e., not restricted to geospatial data), Wang and Strong (1996) identified several characteristics based on a large survey among data users, grouped into four categories: intrinsic (e.g., believability, reputation), contextual (e.g., relevancy, timeliness), representational (e.g., interpretability, ease of understanding) and accessibility (e.g., accessibility, security).

Most of these criteria are not available in today's metadata but would be necessary to help users to assess the fitness for use of datasets for certain applications. For instance, accurate and up to date data may not fit for the intended use if the-data producer is not recognized (reputation), price is extremely high (cost), time to get them is too long (accessibility), or if data sharing is not permitted (legal issues).

Geospatial Data Quality Information Hierarchy

The design of a data model allowing the management of geospatial data quality information requires knowing *how* information about data quality is related to the data being

described. Quality information can, for instance, describe a whole dataset quality or only a subset of it (e.g., quality of the data related to an object class, or quality of the data of a single attribute of an instance). As described by Bédard and Vallière (1995), there are different levels of detail of data quality, also named granularity of data quality. They suggest a method to aggregate quality information from a single data up to the complete dataset. Hunter (2001) identified quality information granularity as one of the major concerns in geospatial data quality research, saying that "data quality suffers generally from being presented at the global level rather than at greatest levels of granularity". Hunter provides several examples illustrating that today's metadata do not provide information at a sufficient level of detail, such as: Positional Accuracy being "Variable", "100 m to 1,000 m", or "±1.5 m (urban) to ±250 m (rural)". The quality of data also varies temporally (e.g., ±30 meters before 1992 to ±10 meters since 1992 for the more recently covered areas), and thematically (e.g., ±$15,000 for residences, to ±100,000 for stores). These examples illustrate that geospatial data quality heterogeneity is not adequately recorded in today's metadata to properly assess data quality for the subset of data being used. A description at a more detailed level would allow for quality information to be provided, such as the positional accuracy of a given road, the precision of commercial value of residences in a given area, or the level of update of building constructions.

Although we are well aware that organizations have difficulties complying with today's metadata standards even for the general dataset level, we believe there exists a need to combine breadth and depth in quality information. The latter can be of varying levels of detail for different features depending on the needs. We also believe, based on Gervais' work (2004), that legal obligations may force data producers and GIS officers to have such detailed information at hand. In fact, this already exists in legally-bound professional activities such as cadastral surveying, property assessment, road building, and other activities where the quality of information is analyzed on a case-by-case basis. Accordingly, this section provides a brief overview of the literature in terms of geospatial metadata levels of detail, looking at metadata standards, academic research, and practical illustrations from the Canadian National Topographic Database (NTDB) metadata.

Some authors suggested hierarchies intended to manage geospatial quality information at different levels of detail (Bédard and Vallière, 1995; Faïz, 1996, 1999; Qiu and Hunter, 1999, 2002).

The ISO 19115 Standard (2003) provides a framework for encoding metadata for the purpose of search and retrieval, metadata exchange, and presentation. This standard proposes a hierarchy that can be used to store metadata at different levels of detail. This hierarchy may assist in filtering or targeting users' queries to the requested level of detail. ISO hierarchy goes further than those of Qiu and Hunter's by allowing the association of metadata to attributes (attribute type and instance).

ISO/TC 211 (2003) metadata levels are:

- *Data series*: A series or collection of spatial data, which share similar characteristics of theme, source date, resolution, and methodology; e.g., a collection of raster map data captured from a common series of paper maps;
- *Dataset*: Consistent spatial data product instance that can be generated or made available by a geospatial data distributor;
- *Feature type*: Spatial constructs known as features are groups of spatial primitives (0-, 1-, and 2-dimensional geometric objects) that have a common identity; e.g., all bridges within a dataset;
- *Feature instance*: Spatial constructs (features) that have a direct correspondence with a real world object; e.g., the Golden Gate Bridge;
- *Attribute type*: Digital parameters that describe a common aspect of grounded spatial primitives (0-, 1-, and 2-dimensional geometric objects); e.g., overhead clearance associated with a bridge;
- *Attribute instance*: Digital parameters that describe an aspect of the feature instance; e.g., the overhead clearance associated with a specific bridge across a road.

Hierarchies can also be identified within metadata provided by data producers. For instance, Canadian National Topographic Database metadata have four explicit levels of detail: dataset, metadata polygon, theme, and geometric primitive, the latest being stored directly in the data file as attributes.

Therefore, several hierarchies were proposed in the literature. If most of them agree on the general levels (e.g., dataset, feature type, and feature instance), they often differ at detailed levels. Indeed, some of them do not address semantic quality (e.g., quality of attributes or semantic values), others do not address geometric primitives values. Regarding the implementation of these hierarchies, some of the approaches are only theoretical while other were tested through prototypes developed using relational databases.

Multidimensional Geospatial Data Quality Management

Juran *et al.* (1974) were the first to define quality as "fitness for use." This definition issued from the quality engineering and management field is now widely recognized in several fields, including the geospatial information community (Chrisman, 1983; Veregin, 1999). ISO 9000 defines quality as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs." We think that quality is not only the "totality of characteristics of an entity," but rather results from similarity measurements between product specifications and user needs. In order to highlight this aspect, we define quality as the closeness of the agreement between data characteristics and explicit or implicit needs of a user for a given application. Quality requires taking user needs into consideration. For this reason, data quality information should not be restricted to the "quality information" section of metadata, but should include further information already available in other sections of metadata standards (e.g., data coverage or spatial reference systems) or information which is not presently available in today's metadata (e.g., accessibility and believability).

Multidimensional Databases: OLAP and SOLAP

In the database field, multidimensional databases, such those used in On-Line Analytical Processing (OLAP), are well-suited for managing information at different levels of detail. Note that the term "multidimensional" is used in this paper according to its definition in the database field and is not restricted to spatial and temporal dimensions (*x, y, z* and *t*). Multidimensional databases are a component of data warehouses designed to support data analyses at strategic and tactical levels of organizations, and they are opposed to the traditional transactional databases that focus on organization transactions. In the context of data warehouse implementation, multidimensional databases do not replace transactional databases, but are complementary by using them as data sources. OLAP systems are tools enabling users to explore and navigate within organizational data structured into a multidimensional database.

OLAP, introduced by Codd (1993), is extensively documented in the database and business intelligence fields. CompInfo (2004) defines OLAP tools as "a category of

software technology that enables analysts, managers, and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user." OLAP tools support both (1) the management of multidimensional data, and (2) the fast retrieval of these data by its users. Their adaptation to the spatial domain, named SOLAP by Bédard (1997), can be found in a small number of papers and books (see Miller and Han, 2001; Rivest *et al.*, 2001) and is emerging today as a powerful complement to GIS (Bédard *et al.*, 2003). Such a SOLAP is being used in this project.

SOLAP tools are good candidates to manage geospatial data quality information because: (a) the heterogeneity inherent to geospatial data, which implies that quality information has to be analyzed and managed at different levels of detail; (b) the need to provide contextual aggregated information is more meaningful to data users. Thus, based on detailed data, SOLAP uses different ways to aggregate different characteristics, themes, regions, and epochs; (c) SOLAP tools offer different techniques of data visualization such as matrices, pie charts, histograms, and maps; (d) SOLAP tools are known to be very fast and easy to use, and they require no knowledge of query languages. SOLAP delivers rapid "keyboardless navigation" through spatial data and spatial operators at different levels of aggregation (Bédard *et al.*, 2003; Marchand *et al.*, 2003); and (e) it appears natural to implement our data quality approach into existing decision-support technologies such as SOLAP because of the spatial heterogeneity inherent to geospatial data and of the increased facility to display and explore quality information (i.e., maps with tables, statistical charts and semantic trees that can be drilled down or up with a single click of the mouse).

OLAP structures are opposed to the traditional On-Line Transactional Processing (OLTP) structures. The latter being classical databases implemented to manage transactions (e.g., bank transactions) as they are oriented towards data processing tasks (entering, storing, updating, integrity checking, securing, and simple querying of data usually at the level of detail they were collected). On the other hand, OLAP systems are oriented towards supporting organizational decision making by providing aggregated data for both present and historical data (Berson and Smith, 1997). OLAP tools rely on multidimensional data models (also called data cubes or hypercubes) that are based on several fundamental concepts such as dimensions, members, measures, and facts. "Dimensions" represent the different themes, or thematic axes, from which a user can analyze the data (thus, differing from the typical *x, y, z* and *t* meaning commonly used in GIS). Dimensions include members organized into hierarchies. Each dimension can have different levels of detail, and each level can include one or several members (i.e., nodes in a tree). For instance, a grocery store can use a dimension "consumer product" including members "vegetable", "salad" and "lettuce" (each member being at a different levels of detail). A "measure" is a piece of information (e.g., total sales) within a fact describing the unique combination of members that make this fact. A "fact" is a unique grouping of instantiated measures defined by the intersection of each dimension (e.g., the fact "$36,000" can be associated to the measure "total sales" for the member "salad" of the dimension "consumer product" when intersected with the member "week 23" of the dimension "time" and the member "Quebec City" of the dimension "region"). Different types of models are possible for the multidimensional database design, such as star and snowflake schemas (Berson and Smith, 1997). Their implementation can be in typical relational DBMS (called ROLAP), in specialized multidimensional databases (called MOLAP) or in hybrid multi-tiers architectures (called HOLAP). The selection of the model depends on the type of data and the expected operations.

Different operators (e.g., drill-down, roll-up, and pivoting) allow users to navigate into the data. For example, the drill-down operator allows navigating in one dimension from a parent member down to a child member, thus getting more details. Roll-up (or drill-up) is the opposite, allowing one to get more global information. These operators do not require any knowledge in database query languages such as SQL (the queries being transparent to the users), and they provide instantaneous answers.

SOLAP tools extended for geospatial data exploration have recently been developed in order to support decision making processes based on geospatial data (Rivest *et al.*, 2001; Bédard *et al.*, 2003). These systems associate OLAP tools with GIS components to enhance geospatial data visualization and analysis. As geospatial data quality may be highly heterogeneous in space, our research aims at integrating the spatial characteristics of data quality into the QIMM model that could be integrated into traditional GIS or SOLAP tools.

Quality Information Management Model (QIMM)

QIMM Dimensions

Information about geospatial data quality (i.e., quality characteristics) can be organized at different levels of detail along dimensions into an OLAP multidimensional database. We suggest in this paper two dimensions that can manage quality information related to most GIS data (Figure 2).

(1) The Quality Indicator Dimension

Quality indicators are a way for data users to get a quick insight on quality information, and then contribute to prevent potential risks (Devillers *et al.*, 2002). Each indicator is based on one or several quality characteristics and is implemented as members of the dimension. In order to avoid information overload, all quality indicators cannot be communicated to data users at the same time. For this reason, they are organized into a hierarchy allowing users to visualize them at different levels of detail. Quality information is aggregated into the dimension hierarchy from the most detailed level to the more general ones. Members of this dimension (i.e., quality indicators) can either provide information regarding spatial (e.g., spatial accuracy), temporal (e.g., temporal accuracy) or thematic (e.g., attribute accuracy) aspects of the dataset. For instance, members can be horizontal positional accuracy, completeness, date of acquisition or accessibility (see Figure 3).

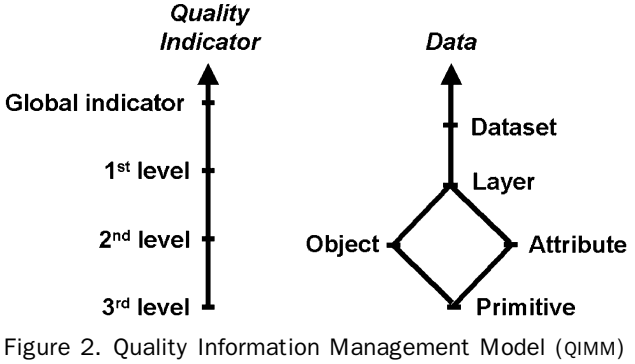


Figure 2. Quality Information Management Model (QIMM) dimensions and members.

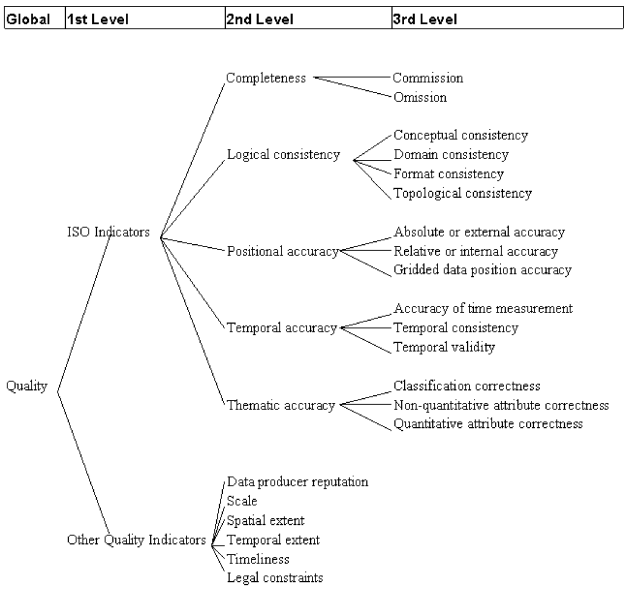


Figure 3. Example of indicators hierarchy. Each indicator is a member of the Quality Indicator Dimension.

Figure 2 presents four levels of detail as an example, but the levels of detail number can be different according to user preferences. The indicator hierarchy does not have to be balanced. For instance, an indicator located on the second level of detail may not be composed of more detailed indicators on the third and fourth levels. Figure 3 provides an example of indicator hierarchy mainly based on ISO/TC 211 19113 and 19115 standards. Users could define their own indicator hierarchy by selecting pre-defined indicators within a database or defining new ones. Global indicator is the more general for a quality indicator, being an aggregation of all first level indicators and providing an insight on the overall data quality. On the other side, the more detailed level is raw quality information issued, for instance, from metadata.

(2) The Data Dimension

The data dimension follows the structure of geospatial data. In this model, quality information is associated to detailed values (e.g., primitive value). Other levels of a dimension hierarchy are either aggregations of the primitive values or raw data if information was only available at more general levels (e.g., average quality of lakes without detailed information about quality of individual lakes). Different aggregation operators available in multidimensional database systems can be used, such as minimum or average or maximum values, depending on user preferences. Other more complex operators can also be implemented and made available to the users (e.g., categorizing, above/under, or quadratic mean square) to support more global analysis of quality information. The Data dimension members are grouped in the following levels:

- Primitive: This level can be either geometric (geometric primitives such as points or lines) or semantic (semantic value). For instance, several geometric primitives can compose an object instance, such as a cadastral parcel composed of several lines (each line being defined by at least two points). As these points can be acquired at different dates or using different technologies, primitives of a same object instance can have different quality levels (e.g., quality related to a point located by GPS or to the value "commercial" of the attribute "type" describing a building);
- Object Instance: This level provides all quality information (geometric and semantic) related to a single instance of object recorded in the dataset (e.g., "Beaver Lake" or "Moose Road"). The overall semantic quality for a certain object is an aggregation of qualities of each data value (e.g., aggregated quality of "Road 138");
- Attribute: This level provides quality related to an object class (or layer) attribute, being an aggregation of primitive value qualities for this attribute (e.g., aggregated quality of attribute "house income" for all buildings instances). Note that only qualities related to semantics can be associated to the attribute level;
- Layer (or Object Class): This level provides the aggregation of the quality (geometric and semantic) of all object instances of a same layer (or class object). A layer can be for instance "roads", "buildings", "rivers" or "parks" (e.g., average quality for all lakes);
- Dataset: Dataset includes quality information (geometric and semantic) related to all objects instances of all data layers. Dataset quality is an aggregation of data layer qualities. Dataset can be, for instance, a topographic map including lakes, rivers, streets, and buildings;

The quality of groups of objects can be aggregated from each object's instance qualities. Such measure can be performed using spatial queries (e.g., what is the overall quality of buildings situated within the city "X" or at less than 500 meters of point "Y"?), or queries on semantics (e.g., what is the overall quality of buildings of "commercial" type or agricultural parcels of "corn" type). In order to benefit from SOLAP performance and ease of use, such groups should be pre-defined.

These levels of the data dimension can include one or several members. Members depend on the datasets manipulated by the users (e.g., members "road" and "river" can become members of the level "layer" when a user adds these data in their GIS environment).

Some intersections between the quality dimensions may be forbidden because of their illogical nature, such as "completeness of a single point" (e.g., fire hydrant) or "positional accuracy of the attribute *building value*."

QIMM Measures

Measures are the piece of information describing quality indicators. Measures should describe both internal (spatial or temporal accuracy, completeness, and logical consistency) and external quality characteristics (difference in currency between a user's expectation and used data, or difference in believability). They can be metadata values or the result of the comparison between metadata values and user needs (e.g., under, equal, or above the needs represented for instance by green, yellow, or red, respectively). As other GIS functions could use quality information stored in the multidimensional database, measures have to be as formalized as possible, avoiding free text for instance, in order to be manipulated more easily by the computer. Quantitative measures are more suitable for data manipulation (e.g., aggregation) than those which are qualitative. Some measures stored in the multidimensional database can be calculated using other measures.

Navigation within the Model and Quality Visualization

Geospatial data users can navigate within the QIMM along both the Data and the Indicators dimensions, moving from a level of detail to another (Figure 4).

For instance (Figure 4a), a user can look at the Global quality indicator for the whole Dataset (position 1: aggregated view of the overall quality for all objects of the dataset). Then, the user can visualize more details along the Data dimension using the OLAP drill-down operator looking at the overall quality for a given layer (e.g., position 2: cadastral parcel layer). Then visualize the overall quality of a single

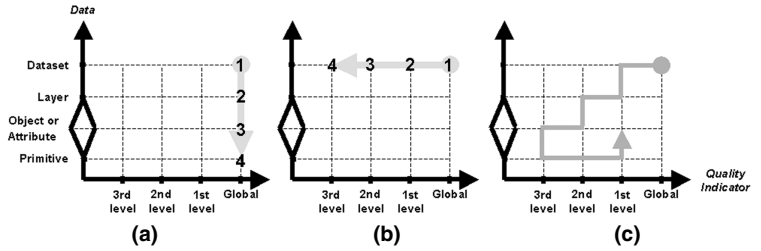


Figure 4. Examples of user navigation into the quality information along both quality dimensions: (a) the Global quality indicator for the whole dataset; (b) quality information along the Quality indicator dimension; (c) an example of complex navigation.

object instance (e.g., position 3: parcel 147), and finally the overall quality of parcel 147's geometric data primitive (e.g., position 4: one of the corners of the parcel). Another navigation scenario (Figure 4b) explores quality information along the Quality Indicator dimension. A user can then start (position 1) at the Global indicator for the whole dataset, then drill-down to the 1st first-level indicator (e.g., position 2: spatial quality), visualizing in this case the average quality related to all objects spatial characteristics. The user can then drill-down to the 2nd level indicator (e.g., position 3: spatial accuracy) still at the dataset level, and finally to the 3rd level indicator (e.g., position 4: horizontal spatial accuracy), being in this case a metadata recommended by ISO and provided into metadata by data providers. Figure 4c provides an example of a more complex navigation, using successive drill-down and roll-up operations along both dimensions. Such navigation allows users to follow its train of thought when exploring quality information provided by a fast and easy user interface such as a SOLAP interface.

Figure 5 provides an example of navigation within quality information displayed in a tabular view using drill-down operations along the two quality dimensions. The first drill-down is performed on the Quality Indicator dimension, allowing the user to move from one level of detail to a more detailed level on this dimension. The second one (i.e., drill-down on Roads) is performed on the Data dimension, allowing the user to move from the "Layer" member down to the "Object" member.

	Positional Accuracy	Completeness	Logical Consistency	Attribute Accuracy
Building	10	100	Good	Low
Hydrology	15	95	Average	Good
Road	10	100	Good	Good
Vegetation	30	90	Good	Average

Drill-down along the Quality Indicator dimension

	Horizontal Positional Accuracy	Vertical Positional Accuracy
Building	5	15
Hydrology	10	20
Road	5	15
Vegetation	15	60

Drill-down along the Data dimension

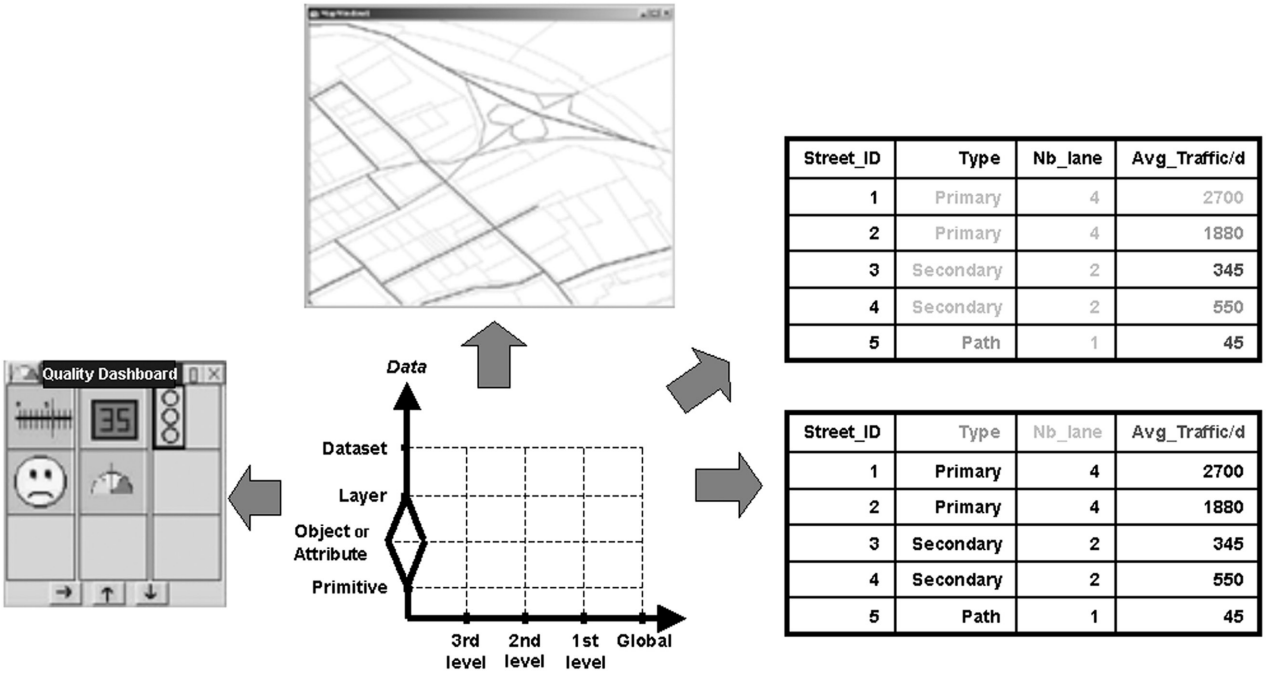
	Positional Accuracy	Completeness	Logical Consistency	Attribute Accuracy
Road 1	10	-	Good	Good
Road 2	15	-	Good	Good
Road 3	5	-	Good	Good
Road 4	10	-	Good	Good
Road 5	10	-	Good	Good

Figure 5. Examples of user navigation in a tabular view using drill-down operator on the two QIMM dimensions.

Based on the QIMM data structure, users can access different displays of quality information, facilitating their analysis. For instance, indicator values can be displayed within a dashboard, on a map, or directly within the descriptive data table (Figure 6). These are examples of possible quality visualization techniques, but a wide range of other techniques can benefit from the quality information stored in the QIMM.

- Dashboard Visualization: Quality indicator values can be displayed within a dashboard (Devillers *et al.*, 2002) such as those used within many decision-support systems. Indicators can have different representations (e.g., number, street light, speed meter, smiley) depending on the type of data to be represented and the user's preference. Figure 6 presents a dashboard including five quality indicators selected by the user for being relevant in his context. Each indicator value is displayed using the representation selected by the user. The dashboard is displayed into the GIS interface and can be visible or not. These indicators represent quantitative or qualitative values resulting from the comparison between data characteristics and users needs. Users can visualize indicators at different levels of detail and are able to navigate into the indicator hierarchy using OLAP operators (e.g., drill-down and roll-up).
- Cartographic Visualization: Indicator values can be displayed on a cartographic base using different representations (e.g., color, shape, texture). SOLAP operators can allow the navigation between the levels of detail in a cartographic view (e.g., moving from the visualization of a quality indicator for a single road to the visualization of the quality of each road segment within this road). This visualization mode is particularly interesting to get an insight on the spatial heterogeneity of quality information, users being able to rapidly identify areas of a map having lower quality and areas having higher quality. Users can also choose the quality parameter they want to visualize (e.g., positional accuracy of objects and temporal accuracy).
- Descriptive Data Table Visualization: Indicators related to semantic quality, such as attribute accuracy or completeness, can be visualized within the data table at different levels of detail. A user can then have a fast insight on the quality of descriptive data within a traditional data table as provided by most GIS software. Figure 6 shows the visualization of values for individual data qualities in the first table (for one instance) and an aggregation of values for data qualities at the attribute level in the second table (i.e., for all instances). This allows users to get a fast insight on the semantic quality of the data manipulated.

The visualization techniques used in a SOLAP (i.e., maps, tables, statistical charts, and semantic tree) allow the users to navigate into quality information from one level of detail



Street_ID	Type	Nb_lane	Avg_Traffic/d
1	Primary	4	2700
2	Primary	4	1880
3	Secondary	2	345
4	Secondary	2	550
5	Path	1	45

Street_ID	Type	Nb_lane	Avg_Traffic/d
1	Primary	4	2700
2	Primary	4	1880
3	Secondary	2	345
4	Secondary	2	550
5	Path	1	45

Figure 6. Possible visualizations of Quality information using the QIMM. Quality information can be for instance displayed in a dashboard (left), on a cartographic base (top), in attribute tables on the individual value level (top right) or on the attribute level (bottom right).

to another along both Indicator and Data dimensions as shown in the next section.

QIMM Prototype

A prototype was developed to test the QIMM model introduced in this paper with a user interface made of a simple dashboard and cartographic visualization. The prototype is based on three main technologies integrated into a single cartographic interface: (1) a multidimensional database storing quality information at different levels of detail into a MOLAP hypercube implemented with Microsoft® SQL Server/Analysis services, (2) cartographic functionalities using GeoMedia® Professional GIS from Intergraph Corporation, and (3) OLAP tools enabling the user to navigate into quality information along the two dimensions of the QIMM model, both in tabular and cartographic views, using Proclarity OLAP software. The resulting SOLAP prototype was tested with data from the Canadian National Topographic Database (NTDB).

This prototype supports functionalities such as:

- Managing quality information into a multidimensional database structure using a subset of the QIMM model (from the data level to the object instance level). QIMM measures are mostly based on quality elements and sub-elements described in the ISO 19113 standard. QIMM dimensions (i.e., data and indicator) were implemented under SQL Server;
- Loading and viewing geospatial data (e.g., zoom in, zoom out, pan, fit all). Spatial objects are linked to the quality information stored in the QIMM using a foreign key;
- Visualizing quality information using indicators displayed into a dashboard and into a cartographic display. Indicators are selected by users within an indicator dataset stored in an Access® relational database.
- OLAP functions (e.g., drill-down, and roll-up) allowing users to navigate into quality information along both data and indicator dimensions.

Quality information issued from metadata is derived into risk levels, based on user-defined tolerance levels. Then, quantitative quality information (e.g., 15 meters for positional accuracy) is compared to a user tolerance level (e.g., 1 meter), and then derived into quantitative values for detailed information or into qualitative values such as green/yellow/red streetlight display for more general information. Qualification of quality information uses user-defined thresholds. Other more complex techniques could be designed as previously mentioned.

Figure 7 shows the main interface of the QIMM prototype. This interface is composed of a cartographic view displaying NTDB dataset, a quality indicator dashboard (located on the left side), and different tools offered to the user. They are from the left to the right: cartographic tools (e.g., pan, zoom in, zoom out, or fit all), QIMM tools (i.e., selection of the quality element to be mapped and definition of the user's tolerance to risk) and some OLAP tools (i.e., drill-down and roll-up). This example shows values for six quality indicators selected by the user (completeness, logical consistency, and up to date) and for a global quality indicator. General quality (aggregation of all quality indicators) was mapped by the user in order to visualize the spatial heterogeneity of quality at the general level.

As seen on Figure 7, an important outcome of this approach is to support the spatial variability of quality information. Indeed, because of the heterogeneity of acquisition methods being used to acquire geospatial data (e.g., total station, GPS, and aerial images), to update them (spatial extent and frequencies and differences in methods), the different objects and geometric primitives contained in a geospatial database can have varying levels of quality. The high level of granularity potentially used for quality information in the QIMM model (down to the geometric or semantic primitives level) allows a very powerful analysis

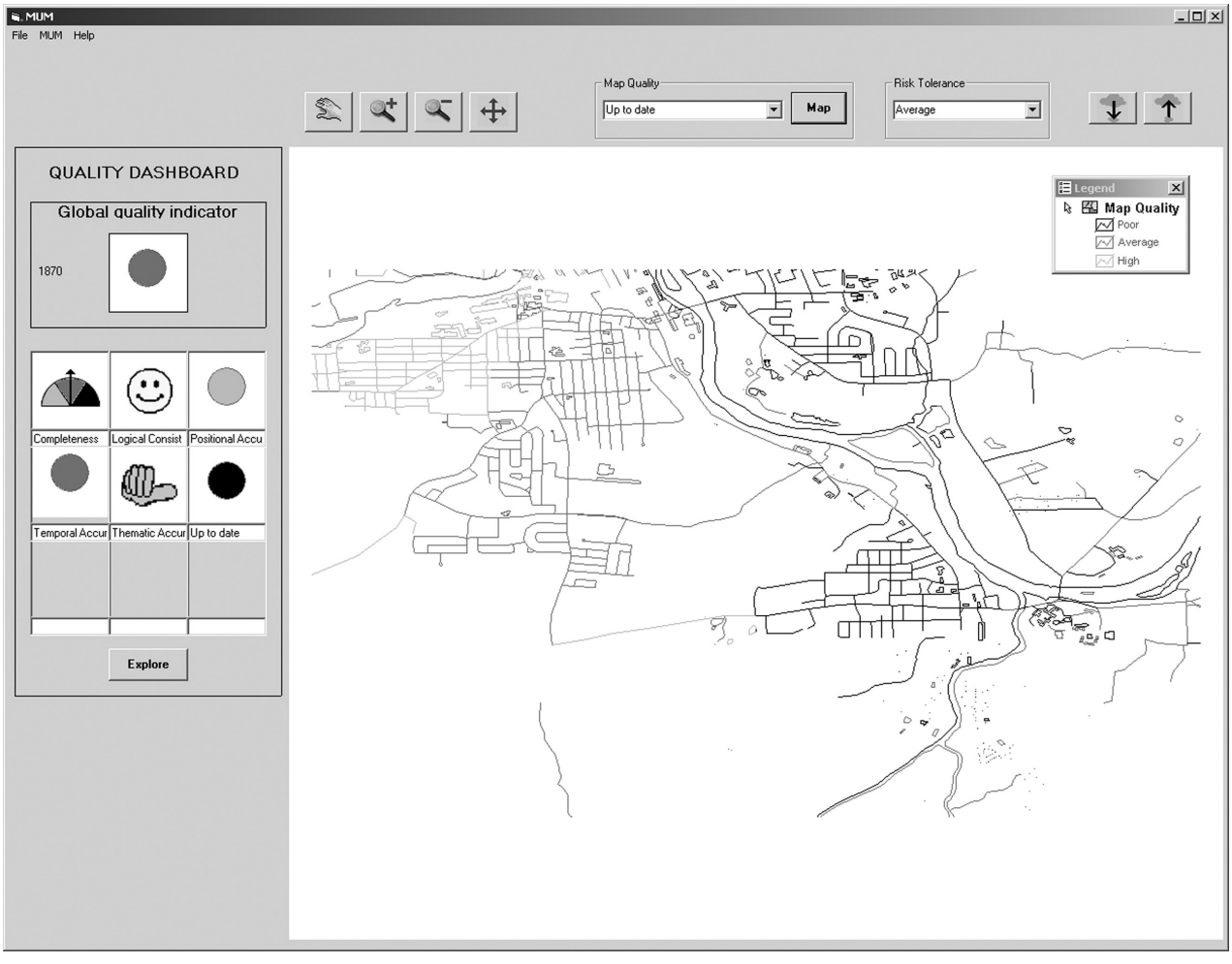


Figure 7. Prototype using the QIMM model to manage and communicate data quality information.

of quality when desired that is the calculation of quality exclusively for the spatial extent defined or visualized by the users. Then, quality information displayed to the user is an aggregation of qualities of every object instances located in the user-defined area or in his cartographic view. Different possible aggregation techniques were mentioned earlier in the paper. Users can then get different quality information (e.g., spatial accuracy, logical consistency, and temporal accuracy) for an area of interest and identify area having higher quality than others. This allows users to get better information on the spatial heterogeneity of quality information.

Conclusions and Perspectives

This paper provides an innovative approach to manage geospatial data quality information based on a multidimensional data management approach. We first highlighted the need to structure quality information in order to provide meaningful and contextual information to geospatial data users. Concepts of Passive, Dynamic and Proactive Multidimensional User Manuals (MUM) were introduced. We presented different works published by standardization and academic bodies classifying data quality into several characteristics. Several works that aimed at recording data quality at different levels of detail were afterwards discussed. Based on the literature, we presented a conceptual framework, named QIMM, allowing the management of quality information at different levels of detail using a multidimensional database approach. QIMM dimensions (i.e., quality indicators and data) and measures were defined and illustrated. Examples of user navigation into quality information were provided to illustrate this approach. Different kind of quality information visualization were presented and discussed. Finally, a prototype based on the QIMM model is presented to test the model and highlight the benefits of such approach in term of diversity of quality information communication.

This work provides a theoretical framework to manage and communicate to users the heterogeneous quality information at different levels of detail. It is rather frequent to find papers mentioning that quality is multidimensional; this work is the first attempt to structure quality information using a multidimensional approach (i.e., SOLAP tools). QIMM provide answers to a main issue of the spatial data quality field: the need to manage various qualities of information at different levels of detail. The model was implemented using commercial multidimensional database, OLAP software and commercial GIS. Such tools can support the users in assessing if the quality of geospatial data is good enough for their needs. In situations where quality information is very heterogeneous and the overall quality

assessment too complex for non-expert users, such tools can help geomatics engineers support non-expert users to assess if the quality is sufficient for their requirements. QIMM implementation is not restricted to multidimensional databases; it is also useful for spatial data quality management in general using traditional relational databases. Quality information being structured at different levels of detail, it can be exploited by different "Quality-aware GIS" programs (e.g., uncertainty management, uncertainty/quality communication and visualization, and error buttons). Furthermore, detailed quality information allows the cartographic visualization of the spatial heterogeneity of quality. Finally, providing aggregated information to users helps to reduce the risk of misuse by reducing the uncertainty they may have regarding data quality. This meta-uncertainty is reduced by both the communication of internal quality information and the communication of risk indicators based on external quality, i.e., the difference between internal quality values and users requirements.

Acknowledgments

This work is part of the MUM project (Multidimensional User Manual) and is funded in part by the Canadian Network of Centres of Excellence GEOIDE, the IST/FET program of the European Community (through the REV!GIS project), the Ministère de la Recherche, de la Science et de la Technologie du Québec, the Centre for Research in Geomatics (CRG) and Université Laval. Thanks are due to Dr. Jean Brodeur, Dr. Gary Hunter, the anonymous reviewers for the critical review of the manuscript, and Geomatics Canada CTI-S for their support.

References

Aalders, H.J.G.L., and J. Morrison, 1998. Spatial Data Quality for GIS, *Geographic Information Research: Trans-Atlantic Perspectives* (M. Craglia and H. Onrud, editors), Taylor & Francis, London/Bristol, pp. 463–475.

Agumya, A., and G.J. Hunter, 1997. Determining the fitness for use of geographic information, *ITC Journal*, 2(1):109–113.

Agumya, A., and G.J. Hunter, 2002. Responding to the consequences of uncertainty in geographical data, *International Journal of Geographical Information Science*, 16(5):405–417.

Beard, M.K., 1989. Use error: the neglected error component, *Proceedings of AUTO-CARTO 9*, March 1989, Baltimore, Maryland, pp. 808–817.

Beard, M.K., 1997. Representations of Data Quality, *Geographic Information Research: Bridging the Atlantic* (M. Craglia, and H. Couclelis, editors), Taylor and Francis, pp. 280–294.

Bédard, Y, 1997. Spatial OLAP 2nd Annual R&D Forum, Geomatics VI, *Canadian Institute of Geomatics*, Montreal, 13–14 November.

Bédard, Y., P. Gosselin, S. Rivest, M.-J. Proulx, M. Nadeau, G. Lebel, and M.-F. Gagnon, 2003. Integrating GIS Components with Knowledge Discovery Technology for Environmental Health Decision Support, *International Journal of Medical Informatics*, 70(1):79–94.

Bédard, Y., and D. Vallière, 1995. Qualité des données à référence spatiale dans un contexte gouvernemental, Technical report for the Ministère des Ressources Naturelles, Université Laval, Québec, Canada (In French).

Berson, A., and S.J. Smith, 1997. *Data Warehousing, Data Mining and OLAP (Data Warehousing/Data Management)*, McGraw-Hill, New York, 612 p.

CEN/TC-287, 1994. *WG 2, Data description: Quality*, Working Paper No. 15, August.

CEN/TC-287, 1995. *PT05, Draft Quality Model for Geographic Information*, Working Paper D3, January.

Chrisman, N.R., 1983. The Role of Quality Information in the Long Term Functioning of a Geographical Information System, *Proceedings of International Symposium on Automated Cartography (Auto Carto 6)*, Ottawa, Canada. pp. 303–321.

Codd, E.F., 1993. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate*, E.F. Codd and Associates.

CompInfo, 2004. CompInfo, The Computer Information Center, URL: http://www.compinfo-center.com/entsys/olap.htm (last date accessed: 18 November 2004).

Curry, M.R., 1998. *Digital places: Living with Geographic Information Technologies*, Routeledge, London & New York, 191 p.

Devillers, R., M. Gervais, Y. Bédard, and R. Jeansoulin, 2002. Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-User Manual, *Proceedings of OEEPE-ISPRS Joint Workshop on Spatial Data Quality*, 20–21 March, Istanbul.

Duckham, M., and J.E. McCreadie, 2002. Error-aware GIS Development, *Spatial Data Quality* (W. Shi, P.F. Fisher, and M.F. Goodchild, editors), Taylor & Francis, London, pp. 63–75.

Elshaw Thrall, S., and G.I. Thrall, 1999. Desktop GIS software. *Geographical Information Systems* (P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, editors), John Wiley & Sons, New York, pp. 331–345.

Faïz, S.O., 1996. *Modélisation, exploitation et visualisation de l'information qualité dans les bases de données géographique*, Ph.D. thesis, Université Paris-Sud, France (In French).

Faïz, S.O., 1999. *Systèmes d'Informations Géographiques: Information Qualité et Data Mining*, Tunis, 362 p. (In French).

FGDC, 2000. *Content Standard for Digital Geospatial Metadata Workbook*, version 2. U.S. Geological Survey, Reston, Virginia.

Frank, A., 1998. Metamodels for Data Quality Description, *Data Quality in Geographic Information – From Error to Uncertainty* (M.F. Goodchild, and R. Jeansoulin, editors), Editions Hermes, pp. 15–29.

Gervais, M., 2004. *La pertinence d'un manuel d'instruction au sein d'une stratégie de gestion de risque juridique découlant de la fourniture de données géographiques numériques*, Ph.D. thesis, Université Laval, Québec, Canada, 344 p. (In French).

Gervais, M., R. Devillers, Y. Bédard, and R. Jeansoulin, 2001. GI Quality and decision making: toward a contextual user manual, *Proceedings of GeoInformation Fusion and Revision Workshop*, 09–12 April, Quebec City, Canada.

Goodchild, M.F., 1995. *Sharing Imperfect Data, Sharing Geographic Information* (H.J. Onsrud, and G. Rushton, editors), Rutgers University Press, New Brunswick, New Jersey, pp. 413–425.

Guptill, S.C., and J.L. Morrison, 1995. *Elements of spatial data quality*, Elsevier Science, New York, 202 p.

Harvey, F., 1998. *Quality Needs More Than Standards. Data Quality in Geographic Information – From Error to Uncertainty* (M.F. Goodchild, and R. Jeansoulin, editors), Hermes Editions, pp. 37–42.

Hunter, G.J., 2001. Spatial Data Quality Revisited, *Proceedings of GeoInfo 2001*, 04–05 October, Rio de Janeiro, Brazil, pp. 1–7.

Hunter, G.J., and K.J. Reinke, 2000. Adapting Spatial Databases to Reduce Information Misuse Through Illogical Operations, *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2000)*, 12–14 July, Amsterdam, The Netherlands, pp. 313–319.

IGN, 1997. Bulletin d'information de l'IGN – Qualité d'une base de données géographique: concepts et terminologie, N. 67, 51 p. (In French).

ISO-TC/211, 2003. *Geographic Information – Metadata19115*.

Juran, J.M., F.M.J. Gryna, and R.S. Bingham, 1974. *Quality Control Handbook*, McGraw-Hill, New York.

Marchand, P., A. Brisebois, Y. Bédard, and G. Edwards, (2004). Implementation and evaluation of a hypercube-based method for spatiotemporal exploration and analysis, *Journal of the International Society of Photogrammetry and Remote Sensing* (theme issue "Advanced techniques for analysis of geo-spatial data"), 59(1-2):6–20.

Miller, H.J., and J. Han, 2001. *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis, 338 p.

Monmonier, M., 1994. A Case Study in the Misuse of GIS: Siting a Low-Level Radioactive Waste Disposal Facility in New York State, *Proceedings of Conference on Law and Information Policy for Spatial Databases*, Tempe, Arizona, pp. 293–303.

Qiu, J., and G.J. Hunter, 1999. Managing Data Quality Information, *Proceedings of International Symposium on Spatial Data Quality*, 18–20 July, Hong Kong, pp. 384–395.

Qiu, J., and G.J. Hunter, 2002. A GIS with the Capacity for Managing Data Quality Information. *Spatial Data Quality* (W. Shi, M.F. Goodchild, and P.F. Fisher, editors), Taylor & Francis, London, pp. 230–250.

Rivest, S., Y. Bédard, and P. Marchand, 2001. Towards Better Support for Spatial Decision Making: Defining the Characteristics of Spatial On-Line Analytical Processing (SOLAP), *Geomatica*, 55(4):539–555.

Timpf, S., M. Raubal, and W. Kuhn, 1996. Experiences with Metadata, *Proceedings of Symposium on Spatial Data Handling, SDH'96*, Advances in GIS Research II, 12–16 August, Delft, The Netherlands, pp. 12B.31–12B.43.

Unwin, D., 1995. Geographical information systems and the problem of error and uncertainty, *Progress in Human Geography*, 19:549–558.

Veregin, H., 1999. Data quality parameters, *Geographical Information Systems* (P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, editors), John Wiley & Sons, Inc., pp. 177–189.

Wang, R.Y., and D.M. Strong, 1996. Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, 12(4):5–34.