5 Data Quality Issues and Geographic Knowledge Discovery

Marc Gervais
Department of Geomatics Sciences, and
Centre for Research in Geomatics, Laval
University, Quebec City, Canada

Yvan Bedard
Department of Geomatics Sciences, and Centre
for Research in Geomatics, and Canada NSERC
Industrial Research Chair in Spatial Database for
Decision Support, Centre for Research in Geomatics,
Laval University, Quebec City, Canada

Marie-Andree Levesque
Department of Geomatics Sciences, and Centre
for Research in Geomatics, and Canada NSERC
Industrial Research Chair in Spatial Database for
Decision Support, Centre for Research in Geomatics,
Laval University, Quebec City, Canada

Eveline Bernier
Department of Geomatics Sciences, and Centre
for Research in Geomatics, and Canada NSERC
Industrial Research Chair in Spatial Database for
Decision Support, Centre for Research in Geomatics,
Laval University, Quebec City, Canada

Rodolphe Devillers
Department of Geography, Memorial University of
Newfoundland, St. John's, Canada and Department
of Geomatics Sciences, and Centre for Research in
Geomatics, Laval University, Quebec City, Canada

CONTENTS

5.1 INTRODUCTION

Geographical data warehouses contain data coming from multiple sources potentially collected at different times and using different techniques. One of the most important concerns about geographical data warehouses is the quality or reliability of the data used for knowledge discovery, decision making, and, finally, action. In fact, this is the ultimate objective aimed by using this type of database. On the other hand, with increasing maturity and the proliferation of data warehouses and related applications (e.g., OLAP, data mining, and dashboards), a recent survey indicated that for the second year in a row, data quality has become the first concern for companies using these technologies (Knightsbridge 2006). Similarly, a recent survey of Canadian decision-makers using spatial data has identified data quality as the third most important obstacle in increasing the use of spatial data (Environics Research Group 2006). Thus, while data quality has become the number one concern for users of non-spatial data warehouses, it is also recognized as an emerging issue for spatial data (Sonnen 2007, Sanderson 2007) and the quality of spatial datacubes is being investigated seriously within university laboratories. In this context, the concept of data quality is making its way into the realm of geographic knowledge discovery, leading us to think in terms of risks for the users, for the developers, and for the suppliers of data, especially in terms of prevention mechanisms and possible legal consequences.

This chapter first introduces the readers to theoretical concepts regarding quality management and risk management in the context of spatial data warehousing and spatial online analytical processing (SOLAP). Then, it identifies possible management mechanisms to improve the prevention of inappropriate usages of data. Based on this theoretical foundation, this chapter then presents a pragmatic approach of quality and risk management to be applied during the various stages of a spatial datacube design and development. This approach aims at identifying and managing in a more rigorous manner the potential risks one may discover during this development process. Such approach has the merit (1) to be applicable in a real context, (2) to be based on recognized quality and risk management models, (3) to take into account lessons previously learned, (4) to encourage proper documentation and, finally, (5) to help clarify the responsibilities for each partner involved in the data warehouse development project. To complete the chapter, associations between these mechanisms and the legal rules governing the relationship between developers and users are presented.

5.2 FUNDAMENTAL CONCEPTS OF SPATIAL DATA QUALITY AND UNCERTAINTY IN A GEOGRAPHIC KNOWLEDGE DISCOVERY CONTEXT

Though data quality has always been an important aspect of geospatial applications, the proliferation of spatial business intelligence (BI) applications in the context of geographic knowledge discovery (GKD) has brought new concerns and raised new issues related to data quality. Their strategic position into organizations is such that these applications may have important impacts on the organization (Ponniah 2001). In order to make informed decisions, decision makers must be aware of the data characteristics and limitations. Otherwise, there is a risk of data misuses or misinterpretations that may cause severe legal, social, and economical impacts on the organization (Devillers et al. 2002). Unfortunately, in the context of GKD and especially with spatial BI applications, several factors increase the risk of data misuses and misinterpretation.

First is the ease with which users interact with the data. As opposed to GIS tools that require specialized knowledge, spatial BI applications are usually based on user interfaces that are easier and do not assume any specific *a priori* knowledge. There is no need to know a query language such as SQL to explore the data or to have specific knowledge about spatial reference methods or internal database structures. By lowering technical skills to operate such applications, they become available to a larger group of users who may not have a complete understanding or knowledge about the spatial, thematic, and temporal characteristics and limitations of the data (Levesque et al. 2007). Also, "the rapidity and ease of data use may lead users to mistakenly feel that data are made-to-order for their decision analysis needs, and hence to deter them from adopting an informed behavior towards data" (Sboui et al. 2008).

Second is the nature of the underlying data warehouses or datacubes. Because GKD applications are often based on data warehousing architectures, the data used have typically undergone several transformations. Building data warehouses or datacubes involves complex data integration and transformation processes (known as ETL procedures, for extract-transform-load) that may affect the meaning of their content (Levesque et al. 2007). Knowing that data sources may also have undergone such processes, it becomes difficult to evaluate the resulting data quality and reliability. Actually, end users of such technologies are rarely aware of these issues, and when they are they rarely receive a robust answer.

Third are the data aggregation methods. GKD and decision makers need aggregated or summarized data to perform their analyses. Hence, aggregation methods must be defined and applied to provide data that will help decision makers and GKD experts to have a global understanding of a phenomenon. This aggregation adds another level of complexity of interpretation (Sboui et al. 2008). Thus, to interpret correctly the data, decision makers must first understand the aggregation method and its impacts on the data.

In short, although spatial BI applications support GKD and the decision-making process, they do not ensure properly informed decisions or quality knowledge. Geospatial data users and decision makers must be aware of data quality in order to reduce the risks of data misuse and misinterpretation (Devillers, Bedard, and Gervais 2004).

5.2.1 Geospatial Data Quality and Uncertainty

In the geospatial literature, the notion of "quality" often mistakenly refers to data precision, uncertainty, or error. Data with good spatial precision are thus often seen as high-quality data. However, the notion of quality goes well beyond the unique concept of spatial precision. In fact, it is usually recognized as including two parts: internal quality and external quality.

Internal quality refers to the respect of data production standards and specification. It is based on the absence of errors in the data and is thus a matter of data producers. According to several standard organizations (such as ISO, ICA, FGDC, and CEN), internal quality is defined using five aspects, also known as the "famous five": (1) positional accuracy, (2) attribute accuracy, (3) temporal accuracy, (4) logical consistency, and (5) completeness (Guptill and Morrison 1995, ISO/TC-211 2002). Information about internal quality is usually communicated to the users using metadata files transmitted with datasets by data producers (Devillers et al. 2007).

External quality evaluates if a dataset is suited for a specific need and hence refers to the notion of "fitness for use" (Juran, Gryna, and Bingham 1974; Chrisman 1983; Veregin 1999; Morrison 1995; Aalders and Morrison 1998, Aalders 2002, Dassonville et al. 2002, Devillers and Jeansoulin 2006). From a user's point of view, a dataset of quality meets or exceeds his expectations (Kahn and Strong 1998). This second definition has reached an official agreement by standardization organizations (e.g., ISO) and international organizations (e.g., IEEE).

AU: Aalders and Morriason, 1998 not in ref list.

Several researchers break down the concept of quality into sub-classes. Veregin (1999), inspired by the work of Berry (1964) and Sinton (1978), defines three components for geospatial data quality: position, time, and theme. He associates these axes to the notion of precision and resolution (spatial, temporal, and thematic precision, etc.). Bedard and Vallière (1995) propose six aspects that can be used to evaluate spatial data quality:

AU: Berry 1964 and Sinton 1978 not in refs.

AU: Bedard andd Valliere 1995 not in refs.

1. Definition is used to evaluate the nature of the data and the object it describes, i.e., the "what" (semantic, spatial, and temporal definitions).
2. Coverage provides information about the space and the time for which the data is defined, i.e., the "where" and "when".

3. Genealogy is related to the data origin, its acquisition methods, and objectives, i.e., the "how" and "why".
4. Precision is used to evaluate the value of a data and if it is acceptable for the expressed need (semantic, temporal, and spatial precision of the object and its attributes).
5. Legitimacy is associated with the official recognition and the legal extent of a data (*de facto* standards, approved specifications, etc.).
6. Accessibility provides information about the facility with which the user can obtain the data (costs, delivery time, confidentiality, copyrights, etc.).

Uncertainty is another inherent aspect of geospatial data and should be taken into account during their exploration and analysis. In fact, any cartographic representation of a phenomenon is an abstraction of the reality according to a specific goal. Given such abstraction and simplification processes, spatial data are, at different levels, inexact, incomplete, and not actual (Devillers 2004). According to Longley et al. (2001), it is impossible to produce a perfect representation of the reality and thus, this representation is inevitably associated with a certain uncertainty. Hence, there is always a risk associated with the use of spatial data that may be inadequate for some decision-making processes.

Bedard (1987) classifies uncertainty into four categories, which combine to provide the global uncertainty associated with an observed reality:

- (1st order) Conceptual, which relates to the fuzziness in the identification of an observed reality
- (2nd order) Descriptive, which relates to the uncertainty associated with the attributes values of an observed reality
- (3rd order) Locational, which relates to the uncertainty associated with the space and time localization of an observed reality
- (4th order) Meta-uncertainty, which relates to the level to which the previous uncertainties are unknown

Though uncertainty cannot be eliminated in spatial databases, mechanisms can be used to (1) reduce it and (2) absorb the residuals (Bedard 1987, Hunter 1999). According to Epstein, Hunter, and Agumya (1998), uncertainty may be reduced by acquiring additional information and improving the data quality. According to Bedard (1987), the residual uncertainty is absorbed when an entity, such as the data producer or the distributor, provides a guarantee for the dataset and will cover potential damages resulting from their use for a given purpose or when the user accepts the potential consequences of using the dataset. Absorption can be shared with insurance companies or by contracting professionals with liability insurance. Uncertainty absorption relates to the monetary risk (e.g.,. in case of damages or a legal pursuit) and makes use of different combinations of the previous means depending on local laws and practices. In all cases, good professional practices and legal liability guidelines require using prevention mechanisms.

5.3 EXISTING APPROACHES TO PREVENT USERS FROM SPATIAL DATA MISUSES

Different mechanisms can be used to improve the prevention of inappropriate usages of spatial data. Existing methods are mostly intended to communicate information regarding data quality, characteristics, and limitations to the users. The traditional method consists of transmitting metadata along with spatial datasets. They are usually provided in separate files and contain highly technical information intended for geographic information system (GIS) specialists. However, such information is too cryptic to be understandable by typical users (Timpf, Raubal, and Werner 1996, Harvey 1998; Boin and Hunter 2007) and one is justified to assume the situation worsens with decision makers or data warehouse users who are further away from the technical details of the data acquisition and ETL processes. Furthermore, metadata are rarely integrated with the data, limiting their consultation and analysis as often required for GKD. In fact, it reduces the possibility of easily exploiting this information directly during the analysis process (Devillers et al. 2007). As an alternative to the actual metadata format, some researchers propose different techniques to communicate data quality information based on different colors, textures, opacities, 3D representations, etc. (McGranaghan 1993, Beard 1997, Drecki 2002, Devillers and Beard 2006). Other researchers propose to provide end users with meaningful warnings when they perform illogical GIS operations (e.g., measure a distance without having first set the geographical reference system) (Beard 1989, Hunter and Reinke 2000). This is related to the concept of error-sensitive or error-aware GIS (Unwin 1995, Duckham 2002).

Other researchers have tackled the fitness for use aspect by improving existing tools to select data that will best fit users' needs (Lassoued, Jeansoulin, and Boucelma 2003), performing risk analysis (Agumya and Hunter 1997), getting opinions from experts (Levesque 2007), and even developing GKD tools to help these experts formulate their opinion by giving them the possibility to integrate, manage, and visualize data quality information at different levels of detail (Devillers 2004, Devillers, Bedard, and Jeansoulin 2005; Devillers et al. 2007, Levesque 2007).

From a data-warehousing point of view, few researchers have tackled the issue of data misuse and misinterpretation. Some have first identified cases where specific online analytical processing (OLAP) operators may lead to inappropriate usages (Lenz and Shoshani 1997, Lenz and Thalheim 2006). Others have suggested restricting the navigation or informing the user when results may be incorrect (Horner, Song, and Chen 2004). Those solutions, however, remain at a theoretical stage and contribute only partially to a global strategy to prevent datacube misuses. They address a subset of the issues related to data warehousing architectures and, above all, they do not consider the spatial aspect of the data. For example, they cannot be used to describe and illustrate the numerous conflicts that must be faced when integrating heterogeneous spatial datasets coming from different producers, or the semantic and geometric aggregations aspects that must be considered for an informed use of datacubes. In fact, most of these solutions are intended for experts in spatial information and data quality rather than the typical users of GKD or BI applications.

AU: Lenz and Thalheim 2006 is listed as 2001 in refs.

5.4 AN APPROACH BASED ON RISK MANAGEMENT TO PREVENT DATA MISUSES IN DATA WAREHOUSING AND GKD CONTEXTS

We suggest using a risk-management approach to face the complexity of the overall data quality issues during the design and feeding of the warehouse datacube. According to ISO/IEC (1999), a risk is defined as a "combination of the probability of occurrence of harm and the severity of that harm." Risk management refers to the reduction of a risk to a level considered acceptable (Morgan 1990, Renn 1998). Our approach is inspired from the risk management approach proposed by ISO/IEC Guide 51 (1999) and considers the notion of "harm" as a data misuse or misinterpretation. Such an approach was proposed by Agumya and Hunter (1999) for transactional geospatial data and is here geared toward multi-themes, multi-scales, and multi-epochs decision-support data underlying GKD applications, and in particular a datacube/SOLAP context. However, the most noticeable difference with the approach proposed by Agumya and Hunter is that the proposed solution takes place during the design process, that is, in a more preventive mode. This key difference relies on the fact that the *raison d'être* and capabilities of datacubes allow us to identify *a priori* the data that will be compared thematically, spatially, temporally, and at different levels of granularity. Several datacubes are typically built from the same data warehouse according to the users' demands and data quality must be analyzed for each application using these cubes. Consequently, the star or snowflake schemas must be designed and populated with data quality in mind to reduce the risks of misuses. As a result, we advocate enriching system development methods (e.g., OMG-MDA or IBM rational unified process) with risk-management processes specific to the prevention of spatial data misuses.

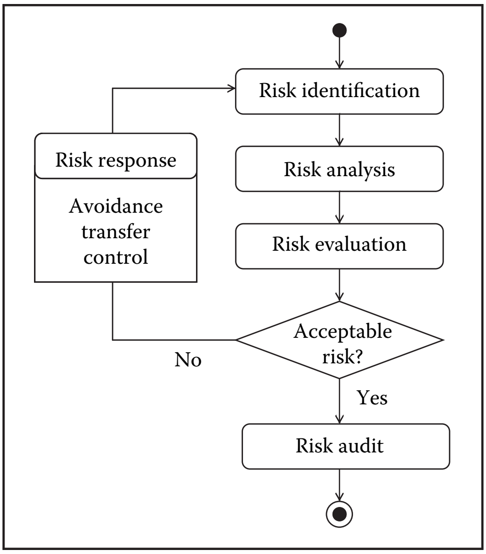
AU: Agumya and Hunter 1999 not in refs.

The proposed approach is a continuous and iterative process that fits with the whole datacube development cycle (needs analysis, design, implementation, feeding). Figure 5.1 shows the different steps proposed: identify risks, analyze them, evaluate potential dangers, prepare responses toward these risks of misuse or misinterpretation, and document the risk-management process as required for quality audits.

AU: Need figure captions for all figures.

Risk identification: This critical step determines the efficiency and quality of the subsequent phases. It aims at finding what could go wrong when using the data, in a way that is as exhaustive as possible. This phase typically involves analyzing (1) the documentation about the data to be integrated (i.e., metadata, data dictionary, source data models), (2) the documentation about the designed datacubes (datacube models, ETL processes, aggregation functions), (3) the material used to train the users, and (4) the existing warnings (e.g., footnotes in reports, tables, and charts, report forewords, restricted accesses, etc.). Once identified, we suggest classifying the risks according to their origin (source) to facilitate further the definition of actions to be undertaken to control them. These categories are: (1) data sources (e.g., missing data), (2) ETL procedures (e.g., erroneous aggregation formula), (3) datacube structure (e.g., not satisfying summarizability integrity constraints), and (4) SOLAP functionalities and operators (e.g., adding on the fly a new measure with faulty formula).

Risk analysis: The second step consists of analyzing, for each risk, its probability of occurrence and the severity of the consequences if it occurs. Risk analysis can rely on different techniques such as simulation techniques or probabilistic analysis.



AU: Caption?

FIGURE 5.1

We can also look at relevant lessons learned during past projects, consult experts and specialists, etc. These two parameters, that is, the probability of occurrence and the severity of the consequences, are usually evaluated according to an ordinal scale composed of three to five levels (e.g., low, moderate, and high).

Like other risk-based approaches, the risk evaluation step is not a simple task; because it demands that we look in a certain way in the future, it often requires experience, judgment, and sometimes intuition. In addition to these, we also consider that an excellent knowledge and understanding of the datacube users' needs and skills, which represent a legal duty of the datacube producer, are necessary to have the best risk analysis possible.

Risk evaluation: The previous results are combined in a matrix to determine the overall level of danger related to each risk (see Table 5.1).

Risk acceptability: Based on the global level of danger previously defined, we must decide whether a risk is acceptable. This analysis is sensitive and should be done very carefully as it may lead to legal consequences. For instance, in front of an acceptable risk (e.g., a risk with an overall level of danger at low), a datacube producer may accept it as it is without communicating the risk to the end-user. In case of damages for the end-user, the data producer can then be legally declared liable to have chosen to ignore it. If the risk is considered unacceptable, the producer must then choose a response mechanism in order to manage it (see risk response in the following).

At this stage, it is recommended to involve end-users and to select with them the appropriate response mechanisms. When users are involved, they understand the risks and approve the mechanisms proposed; they become directly involved in the uncertainty absorption process. Consequently, the datacube designers/providers are better protected in case of problems related to data quality and legal actions.

TABLE 5.1
General Hierarchy Matrix

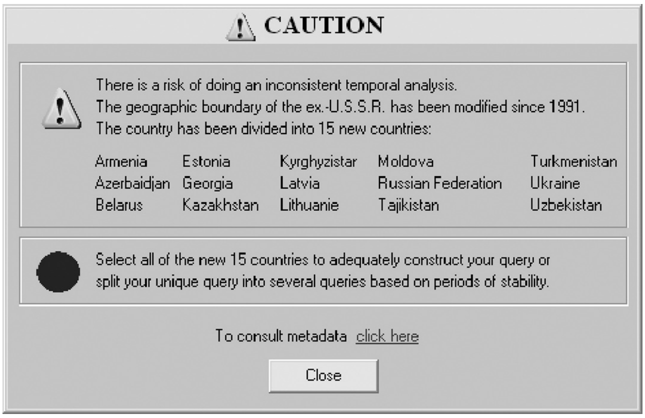
		Probability of Occurrence		
		Low	**Moderate**	**High**
Severity level	High	M	H	H
	Moderate	L	M	H
	Low	L	L	M

Note: L = Low, M = Moderate, H = High.

Adapted from Kerzner, H., 2006. Project Management: A Systems Approach to Planning, Scheduling, and Controlling, 9th ed., John Wiley & Sons, New York.

Risk response: This step is required to manage the risks that are unacceptable to the users. This is where the datacube producer suggests how to cope with those risks. Several mechanisms can be used, such as:

- Avoidance: This mechanism aims at reducing an unacceptable risk by eliminating the source from which it emerges. For example, a data producer may decide not to provide data considered too sensitive or not reliable for the users and their intended usages. Such action is frequent when data are associated with a coefficient of variation above a certain threshold. It is usually applied to moderate to high risks that appear late in the development cycle (Kerzner 2006).
- Transfer: The transfer mechanism is used to move or share a risk with another entity in order to reduce it to a lower level for the datacube producer. For instance, the datacube producer can transfer a risk to a third party (such as an insurance company) who will become liable for the end-user in part or in totality.
- Control: The control mechanism suggests reducing the risk by taking preventive actions. The ISO/IEC Guide 51 standard states that risk reduction must first take place in the design phase, for example by modifying the conceptual model of the datacube or implementing integrity constraints. This is a key step to minimize risks. In addition, Guide 51 suggests producing information for security purposes, that is, “warnings.” General warnings can be communicated to datacube users in a user manual while specific ones (i.e., context-sensitive warnings) can be automatically prompted in the SOLAP application when users are facing a risky query.
- We propose warnings according to the ISO 3864-2 (2004) standard for product safety labels. This standard proposes to communicate (1) the danger level of the risk with standardized alert words (e.g., danger, warning, or caution), (2) the nature of the risk with a symbol, (3) the consequence of the risk, and (4) how to avoid the risk. Figure 5.2 shows such a warning message that could be prompted when analyzing the data.



CAUTION

There is a risk of doing an inconsistent temporal analysis.
The geographic boundary of the ex.-U.S.S.R. has been modified since 1991.
The country has been divided into 15 new countries:

Armenia
Azerbaidjan
Belarus
Estonia
Georgia
Kazakhstan
Kyrghyzistar
Latvia
Lithuanie
Moldova
Russian Federation
Tajikistan
Turkmenistan
Ukraine
Uzbekistan

● Select all of the new 15 countries to adequately construct your query or
split your unique query into several queries based on periods of stability.

To consult metadata click here

AU: Caption?

FIGURE 5.2

The remaining risks must be treated at the end-user level, for example by providing training, limiting access, building user profiles, etc. Defining the category of controls to apply and when to apply them require an excellent collaboration between the datacube producer and the users.

Risk audit: The last step is to document the previous steps as if preparing for an audit. Ideally, this documentation is made while designing and feeding the datacube as it is mainly during these steps that we think about or discover the potential problems. The documentation about a warning must include the message itself, the involved elements, and the triggering elements (e.g., before the SOLAP query or once the results are displayed). This is helped by a series of forms implemented into a UML-based CASE (computer-assisted software engineering) tool, and by a dictionary of terms and definitions describing these processes (Levesque 2008). Such documentation is very important from a legal standpoint because (1) this documentation can help prove the designer/producer complied with their legal duties, (2) it is helpful to prepare training material, and (3) it is an important source of information to manage risks in future datacube developments. More generally, it also helps system designers to build systems that are more robust.

AU: spell out UML.

5.5 LEGAL ISSUES RELATED TO THE DEVELOPMENT OF SPATIAL DATACUBES

Using a risk management approach to prevent data misuse is not only a matter of satisfying users' requirements, it is also a matter of legal liability principles. This section summarizes the legal principles that apply to the datacube producers and datacube users who are linked by a business relationship. First, we describe legal criteria related to the internal quality of data. Second, we describe the criteria related to the external quality when developing a datacube for a given purpose. Third, we summarize the legal duties of datacube users. Finally, we conclude with the pertinence of using a risk-management approach that involves both datacube producers and users during production of the datacube.

5.5.1 Legal Criteria for Spatial Datacube Producers Related To the Internal Quality of Data

Spatial datacubes are a special category of spatial database. They are not yet sold as commercial datasets per se, they are still designed as ad hoc custom services for a given need, and they are populated under the supervision of a professional. In general, it is recognized in many countries (such as in Canada and France) that database production is under the same legal liability regime as information production by agencies (Le Tourneau 2001, 2002; Le Tourneau and Cadiet 2002; Vivant et al. 2002; Dubuisson 2000; Côté et al. 1993). When offering services, the datacube producer must care about internal data quality (quality of the content) and external quality (fitness-for-use and quality of the presentation).

Regarding internal quality, unless specified, database producers are expected to deliver data that are exact, complete, and up-to-date because these are the three most important legal criteria used to assess internal quality (i.e., a subset of the ISO/TC211 data quality indicators). Applying these criteria to spatial datacubes raises some issues (e.g., cascading updates from source databases, completeness of aggregation and summarization of data, exactness of statistical indicators and multi-scale generalized maps, time-varying maps, etc.), especially as spatio-temporal data are known to convey inherent uncertainty that cannot be eliminated (Gervais et al. 2007, Gervais 2004, Bedard 1987). Consequently, as it is the case for databases in general (Lucas 2001), one cannot always expect an internally perfect database as it is often impossible to achieve. Rather, it is expected for the data producer to use appropriate means to achieve the required internal quality. Database producers are thus typically facing an obligation of means as opposed to an obligation of results. Obligation of means refers to the obligation of the provider to act carefully to meet the expectations of the client and consequently to use all reasonable means to achieve the desired result, without warranting a perfect result (Baudouin and Jobin 1998). Consequently, it is expected for a datacube producer to formally adopt procedures especially tailored toward ensuring the internal quality of data, but without imposing the production of perfect data. Legally, the emphasis is given to the verification procedures that are used rather than the result obtained. In particular, in a datacube design or an ETL process, the datacube producer should not perform a task without knowing its impact on the resulting values (e.g., measures in the fact table).

AU: Baudouin and Jobin 1998 not in refs.

5.5.2 Legal Criteria for Spatial Datacubes Producers Related To the External Quality of Data

From a legal standpoint, the external quality is directly related to the diffusion and method of presenting data to the users. When the producer cannot guarantee the exactness of the data (as is typically the case with spatial data), there is an obligation to properly inform the users. Such obligation of proper information is in fact the legal mechanism to deal with imperfect products. It is expected that a producer provide all the information necessary to the users so they can properly assess the adequateness of a product concerning their needs. The level of information to provide is directly proportional to the incompetency of the user and to the level of complexity, technicality, and dangerousness of the product when being used. Consequently, from a legal point of view, evaluating the external quality of a spatial datacube becomes the evaluation of the information delivered with the spatial datacube.

Depending on the level of the uncertainty inherent to the datacube or on the level of dangerousness regarding the use of the datacube, one finds three types of such obligations: typical information, advice, and warning. Typical information does not require influencing the decision of the user (Lefebvre 1998) but it must be provided in a language and level of detail compatible with the expected typical users' level of knowledge. For example, providing only the metadata of a datacube could be sufficient if the user has the necessary knowledge to understand the technical terms related to spatial metadata and their impact on the proper use of the data (e.g., a well-trained and experienced user). The obligation of advising becomes important when the producer estimates that the provided datacube is complex and highly technical, or that the users need specific information because they do not have the necessary background to understand the characteristics of the datacubes or the consequences related to the planned usage (Lucas 2001). Such obligations may lead the datacube producer to perform additional research or analysis or to modify the datacube (Le Tourneau 2002). Finally, the obligation to provide warnings is always there (Baudouin and Deslauriers 1998), especially when one estimates that there are potential dangers to using the datacube. Such warnings must be clearly written, complete and up-to-date, and presented to users as soon as a danger is seen as potential (even before a final conclusion). This is an obligation of prevention that may direct the users away from erroneous usages or toward good usages.

Preventing dangerous usages by providing warnings requires identifying and communicating the anticipated risks (Rousseau 1999). A risk-management approach geared toward users' needs and level of tolerance to risks is mandatory. When uncertainty is high, the datacube producer must increase the degree of awareness of users. Considering the higher level of knowledge of the datacube producer, it is expected that he or she will make up for the users' lack of appropriate knowledge. Several court decisions regarding spatial data support this conclusion (e.g., breaking underground infrastructures,[*] marine charts depth errors,[†] erroneous transportation costs calculation,[‡] airplane crashes with deaths,[§,¶] shipwrecks,[**,††] unreasonable fire truck delay,[‡‡] hunting in the wrong area,[§§] cross-country skier death,[¶¶] and building a house in a forbidden area[***]).

[*] Bell Canada v. Québec (Ville), [1996) A.Q. 172 (C.S.); Excavations Nadeau & Fils. v. Hydro-Québec, [1997) A.Q. 1972 (C.S.).

[†] Fraser Burrard Diving Ltd. v. Lamina Drydock Co. Ltd., [1995) B.C.J. 1830 (B.-C.S.C.).

[‡] Côté v. Consolidated Bathurst, [1990) A.Q. 64 (Qué. C.A.).

[§] Aetna Casualty & Surety Co. v. Jeppesen & Co., 642 F.2d 339 (1981).

[¶] Brocklesby v. United States of America, 767 F.2d. 1288 (9th. Cir., 1985); Times Mirror Co v. Sisk, 593 P.2d. 924 (Ariz.1978).

[**] Algoma Central and Hudson Bay Railway Co. v. Manitoba Pool Elevators Ltd, [1966) S.C.R. 359; Warwick Shipping Ltd. v. Canada [1983) C.F. 807 (C.A.).

[††] Iron Ore Transport Co. v. Canada, [1960) Ex. C.R. 448.

[‡‡] Bayus v. Coquitlam (City), [1993) B.C.C.S. 1751; Bell v. Winnipeg (City), [1993) M.J. 256.

[§§] R. v. Rogue River Outfitters Ltd. [1996) Y.J. 137 (Y.T.C.).

[¶¶] Rudko v. Canada, [1983) C.F. 915 (C.A.).

[***] Sea Farm Canada v. Denton, [1991) B.C.J. 2317 (B.-C.S.C.).

5.5.3 Legal Criteria for Users of Spatial Datacubes

Users of datacubes also have legal obligations to ensure data are properly used. The most important obligations are those of collaboration with the datacube producer, constancy when defining the needs, and consistency when using the datacube (i.e., in accordance to the conditions emitted by the datacube producer) (Le Tourneau 2002). Collaboration must take place continuously when negotiating, defining the expectations, providing the required documentation and information, identifying the potential risks, designing and populating the datacube, and defining the means to deal with the identified risks of usage.

5.5.4 Legal Pertinence of a Risk-Management Approach

From a legal perspective, using a risk-management approach is necessary to protect both the datacube producers and users, in particular:

- Implementing formally such an approach within the datacube development method indicates the producer's will to take the necessary means to control rigorously the development of the cube and the decisions made during this phase.
- Continuously communicating with users allows the datacube producer to better assess their tolerance to risk and to adapt the solutions accordingly. It also increases users' awareness.
- Involving users' collaboration in the complete process helps them to fulfill their duty of collaboration.
- Producing proper documentation helps datacube producers to meet their legal duty for information, advices, and warnings. The documents can be used for users' training or for further reference, and they become tangible proof that the work has been done.

A detailed description of the proposed method is beyond the goal of this chapter; however, it can be found in Levesque (2008). Overall, such an approach helps to clearly share the responsibilities between datacube producers and datacube users with regard to the risks of potential misuses. In addition, it adds rigour in the datacube development cycle, and increased users' satisfaction as well as a higher level of professionalism for the datacube producer.

5.6 CONCLUSION

This chapter focused on spatial datacube quality and, more specifically, on an approach to manage the risks of data misuse. We have synthesized issues related to internal and external data quality and presented how they have impacts on the design, populating, and use of spatial datacubes. This is a very recent concern in the GKD and spatial data warehousing community and indicates a new level of maturity. In particular, we have introduced the basis for adopting a risk management approach while developing datacubes. Such an approach allows reduction of the risks of data misuse, improves the involvement of users in the development of datacubes, and helps identify the responsibilities of the involved participants. Finally, we have made an overview of the legal motivations to adopt such a risk-management approach. Although such an approach cannot prevent all risks of data misuse, it is a mean to prevent such risks and to increase users' awareness, leading to spatial datacubes with higher internal and external quality.

REFERENCES

Aalders, H.J.G.L. 2002. The registration of quality in a GIS, in W. Shi, P. Fisher, and M.F. Goodchild (Eds.), *Spatial Data Quality*, Taylor & Francis, London, pp. 186–199.

Agumya, A. and Hunter, G.J. 1997. Determining the fitness for use of geographic information, *Journal of the International Institute for Aerospace Survey and Earth Science*, 2, 109–113.

Baudouin, J.-L. and Deslauriers, P. 1998. *La Responsabilité Civile*, Les Éditions Yvon Blais Inc., Cowansville.

Beard, K. 1989. Use error: the neglected error component, *Proc. AUTO-CARTO 9*, Baltimore, MD, pp.808–817.

Beard, K. 1997. Representations of data quality, in Craglia, M. and Couclelis, H. (Eds.), *Geographic Information Research: Bridging the Atlantic*, Taylor & Francis, London, pp. 280–294.

Bedard, Y. 1987. Uncertainties in land information systems databases, *Proceedings of Eighth International Symposium on Computer-Assisted Cartography*, Baltimore, MD, pp. 175–184.

Boin, A.T. and Hunter, G.J. 2007. What communicates quality to the spatial data consumer? *5th International Symposium on Spatial Data Quality*, Enschede, The Netherlands, June 13–15.

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.

Côté, R., Jolivet, C., Lebel, G.A., and Beaulieu, B. 1993. La Géomatique, ses enjeux juridiques, Publications du Québec, Québec.

Courtot, H. 1998. La gestion des risques dans les projets, Éditions Économica, Paris, France, pp. 17–74.

Courtot 1998 not cited in text.

Dassonville, L., Vauglin, F., Jakobsson, A., and Luzet, C. 2002. Quality management, data quality and users, metadata for geographical information, in Shi, W., Fisher, P., and Goodchild, M.F. (Eds.), *Spatial Data Quality*, Taylor & Francis, London, pp. 202–215.

Devillers, R., Gervais, M., Jeansoulin, R., and Bedard, Y. 2002. Spatial data quality: From metadata to quality indicators and contextual end-user manual, OEEPE/International Society for Photogrammetry and Remote Sensing (ISPRS) Workshop on Spatial Data Quality Management, March 21–22.

Devillers, R. 2004. Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales. PhD thesis, Sciences Géomatiques, Université Laval, Canada.

Devillers, R., Bedard, Y., and Gervais,M. 2004. Indicateurs de qualité pour réduire les risques de mauvaise utilisation des données géospatiales, *Revue Internationale de Géomatique*, 14(1), 35–57.

Devillers, R., Bedard, Y., Jeansoulin, R. 2005. Multidimensional management of geospatial data quality information for its dynamic use within GIS, *Photogrammetric Engineering & Remote Sensing*, 71(2), 205–215.

Devillers, R. and Jeansoulin, R. (Eds.). 2006. *Fundamentals of Spatial Data Quality,* ISTE, London.

Devillers, R. and Beard, K. 2006. Communication and use of spatial data quality information in GIS, in Devillers, R. and Jeansoulin, R. (Eds.), *Fundamentals of Spatial Data Quality*, ISTE Publishing, London, pp. 237–253.

Devillers, R., Bedard, Y., Jeansoulin, R., and Moulin, B. 2007. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data, *International Journal of Geographical Information Sciences*, 21(3), 261–282.

Drecki, I. 2002. Visualisation of uncertainty, in Shi, W, Fisher, P.F., and Goodchild, M.F. (Eds.). *Geographic Data, Spatial Data Quality*, Taylor & Francis, London, pp.140–159.

Dubuisson, B. 2000. Introduction, dans La responsabilité civile liée à l'information et au conseil, Dubuisson, B. and Jadoul, P., Publications des Facultés Universitaires Saint-Louis, Bruxelles, pp. 9–13.

Duckham, M. 2002. A user-oriented perspective of error-sensitive GIS development, *Transactions in GIS*, 6(2), 179–193.

Environics Research Group. 2006. Sondage auprès des décideurs ayant recours à l'information géographique-2006: Sommaire, Technical report for GeoConnection, Natural Resources Canada, October. Also available in English.

Epstein, E. F. Hunter, G.J., and Agumya, A. 1998. Liability insurance and the use of geographical information, *International Journal of Geographical Information Science*, 12(3), 203–214.

Gervais, M. 2004. Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques, PhD Thesis, Université Laval, Québec, Canada and Université Marne-La-Vallée, France.

Gervais, M., Bedard, Y., Jeansoulin, R., and Cervelle, B. 2007. Obligations juridiques potentielles et modèle du producteur raisonnable, Revue Internationale de Géomatique, Éditions Lavoisier, Paris, 17(1), 33–62.

Guptill, S. C. and Morrison, J. L. 1995. Elements of Spatial Data Quality, Elsevier Science, New York.

Harvey, F. 1998. Quality needs more than standards, in Goodchild, M. and Jeansoulin, R. (Eds.), *Data Quality in Geographic Information — From Error to Uncertainty*, Editions Hermes, pp. 37–42.

AU: Need location of publisher.

Horner, J., Song, II-Y., and Chen, P.P. 2004. An analysis of additivity in OLAP systems, *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP*, Washington, pp. 83–91.

AU: Washington, D.C. or Washington state? If Washington state, need city.

Hunter, G. J. 1999. Managing uncertainty in GIS, in Longley, P. A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W. (Eds.), *Geographical Information Systems, Management Issues and Applications*, John Wiley & Sons, New York, pp. 633–641.

Hunter, G.J. and Reinke, K. 2000. Adapting spatial databases to reduce information misuse through illogical operations, *Proceedings Spatial Accuracy Assessment, Land Information Uncertainty in Natural Resources Management,* Amsterdam, The Netherlands, pp. 313–319.

ISO/IEC Guide 51. 1999. Aspects liés à la sécurité — Principes directeurs pour les inclure dans les normes.

ISO-TC/211. 2002. Geographic Information — Quality principles 19113.

ISO 3864–2. 2004. Safety colours and safety signs — Part 2: Design principles for product safety labels.

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

Kahn, B.K. and Strong, D.M. 1998. *Product and Service Performance Model for Information Quality: An Update, Conference on Information Quality*, Massachusetts Institute of Technolog, Cambridge, MA.

Kerzner, H. 2006. *Project Management: A Systems Approach to Planning, Scheduling, and Controlling*, 9th ed., John Wiley & Sons, New York.

Knightsbridge, 2006. Top 10 Trends in Business Intelligence for 2006. White Paper.

AU: Need more information.

Lassoued, Y., Jeansoulin, R., and Boucelma, O. 2003. Médiateur de qualité dans les systèmes d'information géographique, SETIT International conference (Sciences Electroniques, Technologies de l'Information et des Télécommunications), Sousse, Tunisia.

Lefebvre, B. 1998. La bonne foi dans la formation des contrats, Les Éditions Yvon Blais Inc., Cowansville.

Lenz, H-J. and Shoshani, A. 1997. Summarizability in OLAP and statistical data bases, *Proceedings of the 9th International Conference on Scientific and Statistical Database Management (SSDB)*, Washington, August 11–13, pp. 132–143.

AU: Washington, D.C. or Washington state?

Lenz, H-J. and Thalheim, B. 2001. OLAP databases and aggregation functions, *Proceedings of the 13th International Conference on Scientific and Statistical Database Management (SSDB)*, pp. 91–101.

Le Tourneau, P., 1995. La responsabilité civile professionnelle, Éditions Economica, Paris.

AU: Le Tourneau 1995 not cited in text.

Le Tourneau, P. 2001. Responsabilité des vendeurs et fabricants, Droit de l'entreprise, Les Éditions Dalloz, Paris.

Le Tourneau, P. 2002. Contrats informatiques et électroniques, Dalloz reference, Les Éditions Dalloz, Paris.

Le Tourneau, P. and Cadiet, L. 2002. Droit de la responsabilité et des contrats, Éditions Dalloz, Paris.

Levesque, J. 2007. Évaluation de la qualité des données géospatiales: Approche top-down et gestion de la métaqualité, M.Sc. Thesis, Université Laval, Québec, Canada.

Levesque, M.A. 2008. Approche générique pour une meilleure identification et gestion des risqué d'usages inappropriés des données géodécisionnelles. MSc thesis draft version, Department of Geomatics Sciences, Laval University, Quebec City, Canada.

Levesque, M.-A., Bedard, Y., Gervais, M., and Devillers, R. 2007. Towards managing the risks of data misuse for spatial datacubes, *Proceedings of the 5th International Symposium on Spatial Data Quality,* June 13–15, Enschede, The Netherlands.

Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W. (Eds.). 2001. *Geographical Information Systems and Science*, John Wiley & Sons, New York.

Lucas, A. 2001. Informatique et droit des obligations, dans Droit de l'informatique et de l'Internet, Thémis Droit privé, Presses Universitaires de France, Paris, p. 441–588.

McGranaghan, M. 1993. A cartographic view of spatial data quality, *Cartographica*, 30, 8–19.

Morgan, M.G. 1990. Choosing and managing technology-induced risks, in Glickman, T.S. and Gough, M. (Eds.), *Readings in Risk*, Resources for the Future, Washington, pp. 5–15.

Morrison, J. L. 1995. Spatial data quality, in Guptill, S.C. and Morrison, J.L. (Eds.), *Elements of Spatial Data Quality*, Elsevier Science, New York.

Ponniah, P. 2001. Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals, John Wiley & Sons, New York, pp. 291–312.

Renn, O. 1998. Three decades of risk research: Accomplishments and new challenges, *Journal of Risk Research* 1(1), 49–71.

Rousseau, S. 1999. La responsabilité civile de l'analyste financier pour la transmission d'information fausse ou trompeuse sur le marché secondaire des valeurs mobilières, dans La responsabilité civile des courtiers en valeurs mobilières et des gestionnaires de fortune: aspects nouveaux, Les Éditions Yvon Blais Inc., Cowansville, p. 35–62.

Sanderson, M. 2007. Data quality challenges in 2007, *Directions Magazine*, January 18 issue.

Sboui, T., Bedard, Y., Brodeur, J., and Badard, T. 2008. Risk management for the simultaneous use of spatial datacubes: A semantic interoperability perspective, *Annals of Information Systems*, Special Issue on New Trends in Data Warehousing and Data Analysis, Springer, submitted.

Sonnen, D. 2007. Emerging issues: Spatial data quality, *Directions Magazine,* January 4 issue.

Timpf, S.,Raubal, M., and Werner, K. 1996. Experiences with metadata, in Kraak, M.-J. and Molenaar, M. (Eds.), *Symposium on Spatial Data Handling, SDH'96, Advances in GIS Research II*, Vol. 2 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, in Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W. (Eds.), *Geographical Information Systems*, Wiley, New York, pp. 177–189.

Vivant, M. et al. 2002. Lamy Droit de l'informatique et des réseaux, Lamy S.A., Paris.