A systematic approach for managing the risk related to semantic interoperability between geospatial datacubes

Tarek Sboui[1,2], Mehrdad Salehi[1,2] & Yvan Bédard[1,2]

[1] Department of Geomatic Sciences and Centre for Research in Geomatics, Université Laval, Quebec City, QC, G1K 7P4, Canada
[2] NSERC Industrial Research Chair in Geospatial Databases for Decision-Support
tarek.sboui.1@ulaval.ca, mehrdad.salehi.1@ulaval.ca, yvan.bedard@scg.ulaval.ca

Abstract

Geospatial datacubes are the database backend of novel types of spatiotemporal decision-support systems employed in large organizations. These datacubes extend the datacube concept underlying the field of Business Intelligence (BI) into the realm of geospatial decision-support and geographic knowledge discovery. The interoperability between geospatial datacubes aims at facilitating the reuse of their content. Such interoperability, however, faces risks of data misinterpretation related to the heterogeneity of geospatial datacubes. While, the interoperability of transactional databases has been the subject of several research works, no research dealing with the interoperability of geospatial datacubes has been found. In this paper, we aim to support the semantic interoperability between geospatial datacubes. For that, we propose a categorization of semantic heterogeneity problems that may occur in geospatial datacubes. Then, we propose an approach to deal with the related risks of data misinterpretation. This approach consists of evaluating the fitness-for-use of datacubes models, and a general framework that facilitates making appropriate decisions about such risks. The framework is based on a hierarchical top-down structure going from the most general level to the most detailed level. This paper shows the usefulness of the proposed approach in environmental applications

Introduction

In order to derive the maximum profit from the power of the geospatial data and the efficiency of the datacube structure in the decision making process, geospatial datacubes have been introduced. Geospatial datacubes integrate spatial data with the datacube structure and are recognized as one of the most promising decision-support systems

(Rafanelli, 2003). In some situations, we need the interoperation between geospatial datacubes. The situations where such needs arise are: 1) a simultaneous and rapid navigation through different geospatial datacubes, 2) a rapid insertion of data in a datacube from another one, and 3) an interactive and rapid analysis of phenomena changes by comparing the content of different geospatial datacubes (Sboui et al., 2007).

With the emergence of software agents, semantic interoperability has been viewed as the technical analogue to human communication (Brodeur, 2004; Kuhn, 2005; Sboui et al., 2007). According to this view, each agent tries to interpret the exchanged data as it has been originally intended by another one. However, due to the uncoordinated use of data (i.e., semantic heterogeneity), agents may faulty interpret data or be uncertain about its intended meaning. That is, there is a risk of misinterpreting the exchanged data.

The risks of data misinterpretation are even more pronounced when interoperating geospatial datacubes developed for strategic decision purposes. In fact, strategic decisions made on the basis of inappropriate interpretations of data may lead decision analysts to have inappropriate judgment and to make unwarranted inferences about some aspects of the problem to be solved, and thus to make faulty decisions.

This article aims to support the semantic interoperability between geospatial datacubes. It proposes a risk management approach that allows to identify and asses the related risks of data misinterpretation in a systematic manner based on the quality of datacubes models (i.e., metadata and schema). In the next section, we discuss the risks of data misinterpretation in the interoperability involving geospatial datacubes. Then, we propose a categorization of the semantic heterogeneity that may occur during such interoperability. Then, we propose an approach to identify and evaluate such risks. After that, we propose a method that aims at supporting an intervener (human or software agent) to respond to these risks. Then, we provide an example of application, and we show the usefulness of the proposed approach in environmental applications. Finally, we conclude this paper.

Risk of data misinterpretation related to semantic interoperability between geospatial datacubes

This section discusses the risks of data misinterpretation related to semantic interoperability in general and the one involving geospatial datacubes in particular.

Interoperability between geospatial datacubes

A datacube is composed of a set of measures aggregated according to a set of dimensions with different levels of granularity. Both dimensions and measures of a geospatial datacube may contain geospatial components (Bédard et al., 2001). Interoperating geospatial datacubes may involve one or the combination of the following actions on their components: 1) integrating measures which may refer to adding a new measure to a datacube from another one based on common dimensions and members, or creating a new common measure based on existing measures of different datacubes, 2) integrating dimensions which may refer to creating a new dimension based on the dimensions of different datacubes, adding one or several dimensions of one datacube to another, or modifying a dimension of a datacube by using existing dimension's levels of another datacube, or 3) comparing a dimension or a measure against another.

Overview of the risks of data misinterpretation in geospatial datacubes interoperability

With the emergence of software agents, semantic interoperability has been viewed as the technical analogue to human communication (Brodeur, 2004; Kuhn, 2005; Sboui et al., 2007). According to this view, each agent tries to interpret the exchanged data by comparing them with his/her knowledgebase content (i.e., ontology). In order for the interoperability process to work properly, the receiver should interpret data as it was originally intended. However, this is not always the case, since agents (the sender and the receiver) may use different data to represent the same real-world phenomena. Also, agents may use same signs to represent different real-world phenomena, depending on the context. This uncoordinated use of data among agents may cause a risk of misinterpreting. That is, the receiver may faulty interpret data or be uncertain about its intended meaning. The uncoordinated use of data is known as the semantic heterogeneity which is know as the

major barrier for the semantic interoperability (Bishr, 1998; Brodeur, 2004; Kuhn, 2005; Staub et al., 2008).

For example, as shown in Figure 1, due to type of cartographic generalization being carried out, the same real-world phenomena (i.e., a building block) may be represented differently (Figure 1 (a and b)), and therefore, data may be interpreted differently. Figure 1 (a) is interpreted as a block with 11 buildings, while Figure 1 (a) is interpreted as a block with 8 buildings. Consequently, there is a risk of misinterpreting the Figure 1 (b) due to the cartographic generalization process.

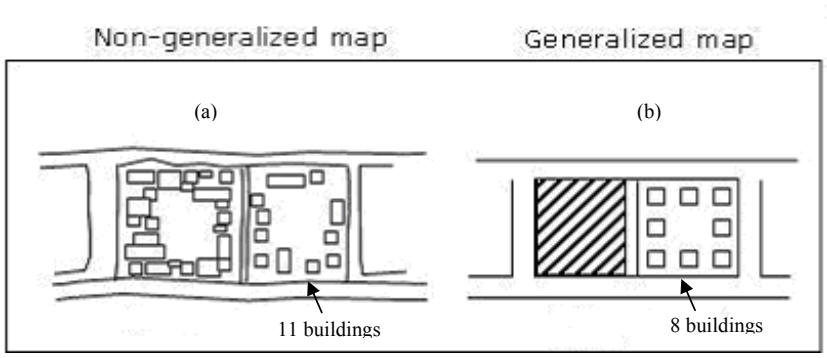

Figure 1. Example of the spatial aggregation-generalization mismatch (Bédard et al., 2005).

Data misinterpretation may lead to a poor understanding of real-world phenomena which may harm the reuse of geospatial data. The risks of data misinterpretation is even more pronounced in the context of the interoperability between geospatial datacubes as they may cause inappropriate judgments, and unwarranted inferences about some aspects of the problem to be solved, resulting in bad strategic decisions.

Furthermore, dealing with the risk of data misinterpretation in geospatial datacubes is more complicated than in transactional databases as geospatial datacubes undergo the complex ETL (Extract, Transform and Load) procedures that can impact the meaning of their content. Some interpretations may be formed, several rules may be applied to the collected data in order to fit the business needs, and different methods may be used to aggregate data. This adds another level of complexity of interpretation as users may need to understand the method or rules used in the ETL procedures.

Semantic heterogeneity conflicts in geospatial datacubes: The main cause of data misinterpretation risk

As for transactional databases, semantic heterogeneity in geospatial datacubes occurs when there are differences in schemas and metadata. Schema heterogeneity refers to the difference in structure of datacube elements (e.g., hierarchy structure) such as the difference between the number of levels of semantically related[1] dimensions of two different geospatial datacubes. For instance, the hierarchy of dimension *Administrative Region* can have the following levels in one datacube: *Country, Province, County, Municipality*; and the following levels in the other one: *Country, Region, Municipality*. Metadata heterogeneity refers to the difference of all other elements of the model of a datacube such as the difference of the precision of the cartographic localization of the geospatial members.

In order to better illustrate our categorisation approach, we present a running example of using two geospatial datacubes (see Figure 2). To determine the risk of forest fire on population, we can interoperate two geospatial datacubes C_1 and C_2, modelled respectively in figure 2 (a) and figure 2 (b). The first datacube C_1 is used to determine the distribution of the population in specific areas and periods. The second geospatial datacube C_2 is used to control the forest fire extent. The example is extracted from a real case which consists in determining the risk of forest fire on the Canadian population.

The geometric geospatial levels as well as geometric geospatial measures are represented using geospatial pictograms which refer to the geometric primitives in the geospatial data modeling (Bédard & Larrivée, 2008). In our example, we use the pictograms developed in Perceptory[2] tool where the pictogram «⊡» represents a point type, the pictogram «☑» represents a line type, and the pictogram «▣» represents a polygon type.

[1] Concepts having a similar representation (e.g., synonyms), generalization, or specialization relationship.
[2] *Perceptory*'s Web Site: http://sirs.scg.ulaval.ca/perceptory

In this example, the metadata of the geospatial datacubes contain information related to the geospatial referencing system, scale and precision, year of creation, stand forest method as well as the geospatial cover of some dimensions.

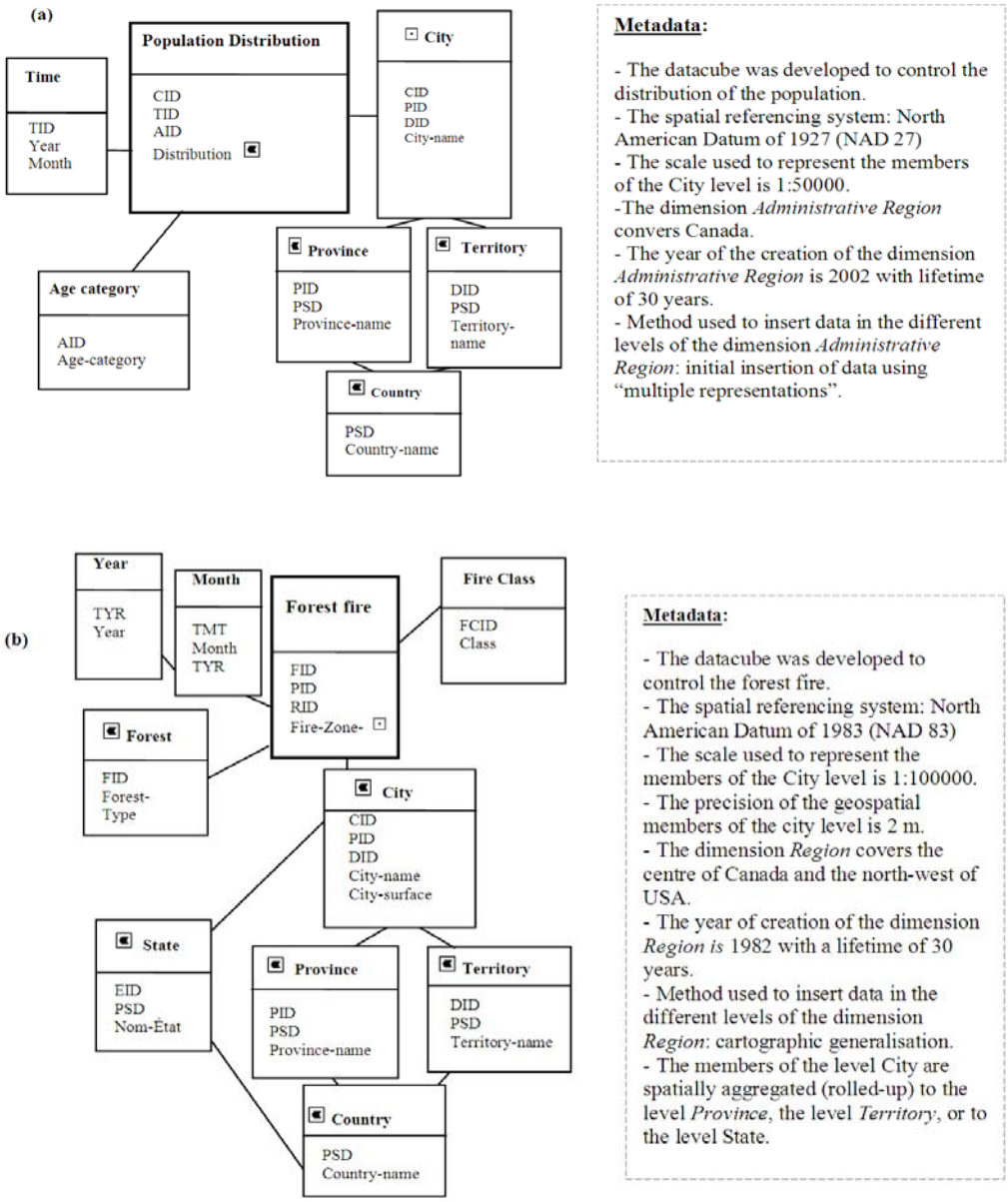

Figure 2 (a and b). Two models of datacubes (C_1 and C_2).

The problems of heterogeneity of geospatial datacube models can exist at five different levels: datacubes, dimensions, hierarchy, levels, and measures. At each level, the heterogeneity can be occurred at the schema and metadata. In order to categorise the problems of heterogeneity, we formalize the elements of datacubes as below:

1. A level $L = \{a_1, a_2,\ldots, a_n\}$ is a set of *n* attributes where *n* is called the cardinality of the set, for example, the level *City* = {city_name, city_surface}.

2. A hierarchy $H = (\{L_1, L_2,\ldots, L_h\}, <)$ is a set of *h* levels and an order relation between these levels. *h* is called the cardinality of the set. The order relation is defined as follows: $\forall\ L_1, L_2 \in H, L_1 < L_2$ if L_1 rolls-up to L_2, for example, the hierarchy (*City*, *Province*, *Country*) of the dimension *Administrative Region* = (*City* < *Province* < *Country*).

3. A dimension $D = \{H_1, H_2,\ldots, H_d\}$ is a set of *d* hierarchy. *d* is called the cardinality of the set and is usually equal to 1, for example, the cardinality of the dimension *Region* = {(*City* < *Province* < *Country*)} is 1.

4. A measure *m* is an attribute which describes the subject of analysis.

5. A datacube $C = (\{D_1, D_2,\ldots, D_c\}, \{m_1, m_2,\ldots, m_q\})$ is a set of *c* dimensions and *q* measures.

The following are the five categories of heterogeneities that may occur between geospatial datacube models:

1. Cube-to-Cube heterogeneity

    a. *Schema Heterogeneity*. It appears when two datacubes have semantically related measures which are represented according to various numbers of dimensions. For example, the two measures *road intersection* and *points of intersection* can be analyzed according to two dimensions in one datacube, and to three dimensions in another.

    b. *Metadata Heterogeneity*:

- *Difference of the creation date of datacube*. It appears when two datacubes were created in two different dates.

2. Measure-to-Measure Heterogeneity

   a. *Schema Heterogeneity:*

      - *Difference in geometric primitives*. It appears when two semantically related measures have different geometric primitives (e.g. line versus polygon).

   b. *Metadata Heterogeneity:*

      - *Heterogeneity of aggregation.* Appears when different functions were used to aggregate semantically related geospatial measures. For example, the functions "geometric union" and "centre of gravity" are used to aggregate the measure *density* in two different datacubes.

3. Dimension-to-Dimension Heterogeneity

   a. *Schema Heterogeneity:*

      - *Inequality of the number of hierarchies.* It appears when the cardinalities of semantically related dimensions are different. Let us suppose that n_1 is the cardinality of dimension D_1, n_2 is the cardinality of dimension D_2. If $n_1 \neq n_2$ then there is a heterogeneity Dimension-to-Dimension at the schema level. In Figure 2, the dimension *Administrative Region* has only one hierarchy: (Country, Province, and City), whereas the dimension *Region* has two hierarchies: (Country, Province, Territory, and City) and (Country, State, and City).

   b. *Metadata Heterogeneity:*

- *Not-correspondence of the dimensions constraints*[3]. It appears when constraints of two semantically related dimensions are incoherent. For example, the constraint of dimension *Administrative Region* indicates that all the members of the level *City* roll-up to the level *Province*, while the constraint of dimension *Region* indicates that the members of the level *City* roll-up to level *Province*, to the level *Territory*, or to the level *State*.

4. Hierarchy-to-Hierarchy Heterogeneity

   a. Schema Heterogeneity:

   - *Inequality of the number of levels.* It appears when the cardinalities of the hierarchies of semantically related dimensions are unequal. Let n_1 be the cardinality of the hierarchy H_1 and n_2 the cardinality of the hierarchy H_2. If $n_1 \neq n_2$, then there is a schema heterogeneity for two hierarchies. For example, in Figure 2, the geospatial hierarchy (Country, Province, Territory and City) of the dimension *Administrative Region* contains four levels, whereas the geospatial hierarchy (Country, State, and City) of dimension *Region* contains three levels.

   - *Inequality of order of levels.* It occurs when hierarchies of semantically related dimensions have different orders of levels. More precisely, for each combination of couples of semantically related levels $((n_1, n_2), (n_1', n_2'))$, if $n_1 < n_2$ and $\neg (n_1' < n_2')$, then there is a hierarchy heterogeneity (inequality of order of levels). For example, in a given datacube, the order of the levels of a geospatial hierarchy is: *City, County,* and *Province*. Whereas, in another datacube, the same levels have the following orders: 1) *City, County,* and *Province*, and 2) *City* and *Province*.

[3] Hurtado et al. (2005) introduced the notion dimension constraint.

b. Metadata Heterogeneity:

- *Heterogeneity of geospatial coverage*. It appears when the members of the hierarchies of two semantically related dimensions have different territorial coverage. For example, the hierarchy of dimension *Administrative Region* (Country, Province, Territory and City) covers Canada, while the hierarchy of dimension *Region* (Country, State and City) covers the North-West of the United States.

5. Level-to-Level Heterogeneity

a. Schema Heterogeneity:

- *Difference in geometric primitives.* It arises when, in two geospatial datacubes, two semantically related levels have different geometric primitives. For example, in datacube C_1, the level City is represented with a point, whereas in datacube C_2, the same level City is represented with a polygon.

b. Metadata Heterogeneity:

- *Difference in geospatial referencing system.* It occurs when there is a difference of geospatial referencing systems used for the members of semantically related levels. In example of Figure 2, the levels of the dimension *Administrative Region* are based on the North American Datum 1927, while those of dimension *Region* are based on the North American Datum 1983.

- *Heterogeneity of cartographic scale.* It appears when semantically related levels are represented with different cartographic scales. In our example, the cartographic scale of the level Province of datacube C_1 is 1:50000, whereas the scale of the level Province of the datacube C_2 is 1:100000.

Only some examples were mentioned in the above categorization but the aspects of heterogeneity can include many others, such as geospatial positioning methods in referencing systems, techniques of data acquisition, algorithms of data transformation, cartographic generalization, and so on.

These problems of heterogeneity may cause risk of misinterpreting data in the semantic interoperability between geospatial datacubes. This paper aims to manage such risks. In the following sections, we propose a risk management approach to manage such risks. The approach consists of evaluating the quality of the conceptual models of datacubes involved in the interoperability process to identify and evaluate the risks of data misinterpretation, and of framework to facilitate making decisions to respond to such risks. The proposed framework corresponds to the proposed categorization of semantic heterogeneity suggesting to respond to the risks at one category at a time (from the more general level to the most detailed level).

An approach to identify and assess the risk of data misinterpretation: model quality perspective

Phases of managing the risks of data misinterpretation

Risk management paradigm constitutes an established framework for evaluating and managing malfunction or potential hazard that may cause bad consequences (Renn, 1998). Inspired from risk management paradigm in the project management literature, we define four iterative phases to manage the risk related to data misinterpretation (see Figure 3): identifying the risks, assessing the risks (i.e., determining their probability of occurrence and their degree of harm), and responding to the risks by making appropriate decisions related to the semantic interoperability process. The possible responses to risk include: reduction, absorption, and transfer (Agumya and Hunter, 2002; Bédard, 1986). We should note that, in the last phase, the risk may be ignored and users will assume the consequences of such choice.

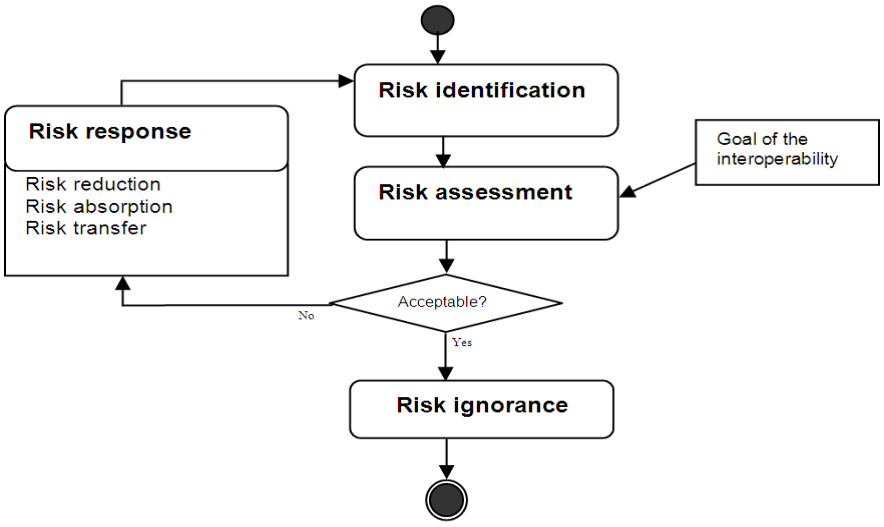

Figure 3. Phases for managing the risk of data misinterpretation

Risk identification and assessment are essential steps to determine how to manage risks. In order to identify and evaluate the risks of data misinterpretation, we measure the external quality (fitness-for-use) of the conceptual models of geospatial datacubes, i.e., metadata and schema.

Conceptual model quality: An aspect to identify and evaluate the risks of data misinterpretation

Conceptual models are central to information systems design and are used as a basis for developing information systems to meet user's requirements at different levels (Moody, 2005).

While a good fitness-for-use of conceptual model (metadata and schema) indicates it is less likely to have a risk of data misinterpretation, a poor quality indicates a higher risk. Consequently, the fitness-for-use of conceptual model allows to identify and evaluate the risk of data misinterpretation. Moreover, evaluating the fitness-for-use of conceptual models facilitates the response to the risks of data misinterpretation as it facilitates the comparison of exchanged data. Indeed, based on such fitness-for-use interveners (software agents or humans) involved in semantic interoperability can be advised:

a. not to use one of the heterogeneous elements of the conceptual models (the one which has a lower quality compared to the other).

b. to consider a model element which has an excellent quality.

c. to use a model element which has a good quality to create new one.

d. not to consider two heterogeneous elements of two conceptual models if they both have a poor quality (that is likely to produce a poor result).

Consequently, our method to identify, evaluate and respond the risks of data misinterpretation is based on evaluating the fitness-for-use of conceptual model: the fitness-for-use of metadata (e.g., coverage of geospatial data) and the one of geospatial datacube schema (e.g., geometric primitives, the number of levels).

The subject of conceptual model quality evaluation occupies a substantial part of the effort devoted to conceptual modelling (Genero et al., 2007). This subject received further emphasis with the Model Driven Development (MDD) paradigm in which development effort is focused on the design of models, rather than on coding (Genero et al., 2007). However, research works focused more on schema quality than the quality of the metadata. Moreover, while a range of quality frameworks have been proposed in the literature, none of these have been widely accepted in practice and none has emerged as a potential standard. As a result, conceptual models have been evaluated in an *ad-hoc* manner (Moody, 2005).

The quality of schema

The impact of conceptual schema quality is of central concern to computer scientists, as well as to end-users (Cherfi et al., 2007). However, while there is a little agreement among experts as to what makes a ‘‘good’’ schema, there are neither guidelines nor standards for evaluating the quality of conceptual schema. Consequently, people continue to evaluate conceptual schemas in an *ad-hoc* and subjective manner based on common sense and experience (Moody, 2005).

The quality of metadata

While a good metadata quality indicates it is less likely to have a risk of misinterpreting data, a poor quality indicates a higher risk, and may undermine the reuse of geospatial data (Agumya & Hunter, 2002), i.e., the main aim of semantic interoperability. Consequently, evaluating the quality of metadata would help to identify and assess the risk of data misinterpretation. While significant research efforts have been carried out to evaluate and enhance the quality of geospatial data (e.g., Agumya & Hunter, 2002; Boin, 2008; Devillers et al., 2007; Frank, 2007), there has been no work on the fitness-for-use of geospatial metadata.

Indicators to identify and evaluate the risks of data misinterpretation

In this section, we define a restricted set of indicators and a method for evaluating the fitness-for-use of metadata and geospatial datacube schema. The definition and evaluation of theses indicators takes into account data requirements of end-users which are represented in the form of a geospatial datacube model (see Figure 4).

We define two categories of indicators: a category related to the schema (*relevance of the geometric primitive*, *relevance of the structure* and *relevance of the hierarchy order*) and another category related to the metadata (*relevance of the metadata* and *freshness of metadata*). Each indicator is evaluated according to a function. The resulting quality value is within the interval (0, 1). The value 1 indicates perfect quality, and hence a low risk. The value 0 indicates completely poor quality, and hence a higher risk. This value is defined in a pragmatic and mathematical way as explained in the following paragraphs. Although quantitative values are used, the quality value represents a scale of ordinal measure. Consequently, we can apply the operators =, >, < to compare the fitness-for-use of elements of different metadata or different schemas. In other words, a quality 0.8 is passably better than a quality 0.4, but it is not precisely twice better. By using such ordinal scale of measures, we are avoiding to attribute a “too quantitative” valued to the result provided to the interveners. We should notice that the proposed indicators do not aim at being complete or precise but rather at making agents globally aware of the fitness-for-use of geospatial datacube models.

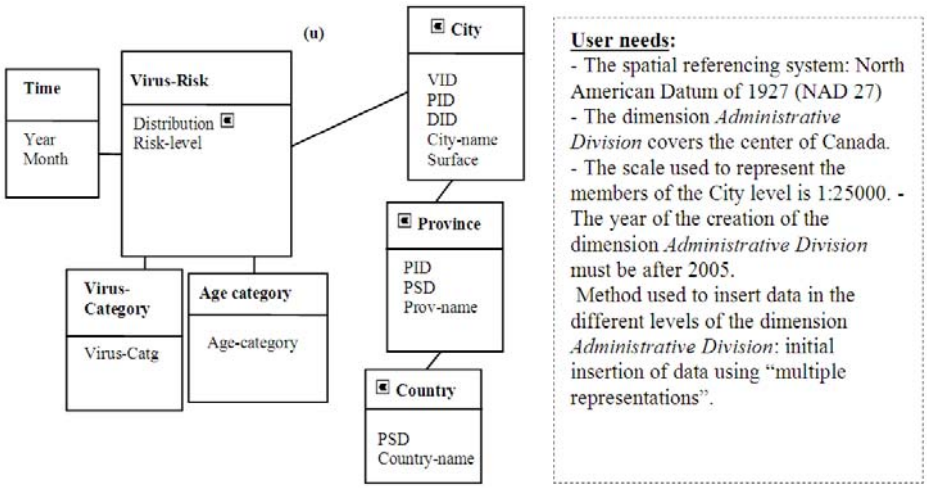

Figure 4. Geospatial datacube model describing the needs of the end-user.

Indicators for evaluating the fitness-for-use of schema elements

For evaluating the quality of the schema, we define the following three indicators:

1- Relevance of the geometric primitive: this indicator evaluates the relevance of the geometric primitive used to represent an element of the datacube model compared to the geometric primitive required by end-users. The indicator can be evaluated based on predefined correspondences between the two primitives. Table 1 represents an example for defining the degree of correspondence between these primitives. The values of this table are defined based on the rules of cartographic generalization and the common practice of cartographers. We set the values 0, 0.25, 0.5, 0.75, and 1 respectively when the primitive of the model: 1) is not sufficient to generalize the required primitive (e.g., a point should not be used to generalize a line), 2) can be used to generalize the required primitives, but the result of generalization will be a bad representation of reality (e.g., a point (0D) and a line (3D)), 3) can be used to generalize the required primitives, but the result of generalization will not faithfully reflect reality (e.g. a point (0D) and a line (1D)), 4) are sufficient to generalize the required primitives, this result will faithfully reflect reality (e.g. a line (1D) and a point (0D)), and 5)is similar to those required (e.g., a line is relevant to represent a line). Geometric primitives are defined for the levels or measures of a geospatial datacube.

Comment [f1]: is (?)
ou bien « the primitives of the mode : » (?)

	Elements of the required model				
Element of the models to be integrated		0D	1D	2D	3D
	0D	1	0	0.5	0.75
	1D	0	1	0	0
	2D	0.75	0	1	0.5
	3D	0.5	0	0.75	1

Table. 1 - Evaluation of the relevance of the geometric primitive.

2- Relevance of the structure: this indicator evaluates the relevance of an element of the geospatial datacube schema with regard to an element of the required schema. It indicates the ratio of the schema elements which are semantically related to those of the required schema. Examples of relevance of the structure are relevance of the number of hierarchies, relevance number of levels and relevance of the number of attributes. The relevance of the structure P_S is calculated as follow:

$$P_s = \begin{cases} \frac{N_{CE}}{N_{CR}} \; ; if \quad N_{CE} < N_{CR} \\ 1 \qquad ; Otherwise \end{cases} \tag{1}$$

Where NR_{CR} is the total number of components required by the user, and NR_{CE} is the number of components which are semantically related to the required components.

This indicator is evaluated for all elements of geospatial datacubes (cube, measure, dimension, hierarchy, and level).

3- Relevance of the hierarchy order: it indicates the relevance of the order of each heterogeneous hierarchy with regard to the structure of the required hierarchy. Suppose that n_1 and n_2 are two levels belonging to the hierarchy H of a given datacube and n_1' and n_2' are two levels belonging to the hierarchy B of the datacube expressing the user's needs (i.e., needs for interoperability). The relevance of the hierarchy order (O_s) is the average of all

the elementary relevance between each pair of levels which belong to the hierarchy of each datacube (o_s). The elementary relevance o_s is determined using the following expression:

$$\forall n_1,n_2 \in H\ ,\ \forall n_1^{'},n_2^{'} \in B:\left[sem\_rel(n_1,n_1^{'}) \wedge sem\_rel(n_2,n_2^{'}) \wedge\ n_1 < n_2 \right] \Rightarrow n_1^{'} < n_2^{'} \qquad (2)$$

Where *sem_rel* is a function that verifies if the levels are semantically related or not. Value 1 (or 0) is set to the elementary relevance (o_s) when the expression (2) is true (or false). Thus, if for example the structure quality of the element 1 is greater than that of the element 2, we can conclude that the structure of the element 1 is more relevant than that of the element 2. As its name indicates, the indicator *relevance of the hierarchy order* is evaluated only for the hierarchies of a geospatial datacube.

Indicators for evaluating the fitness-for-use of metadata

The fitness-for-use of metadata refers to the relevance of the metadata that should be considered for a specific application. We propose a set of indicators and a quantitative approach to evaluate the fitness-for-use of the geospatial datacubes metadata. The following two indicators are proposed:

1- Relevance of metadata. It indicates the degree of relevance of metadata with regard to the needs of end-users. The relevance is evaluated at various levels: thematic, geospatial and temporal. It is evaluated based on the ratio of the number of metadata elements, which are semantically related to the required elements, to the total number of required elements. Relevance of the metadata *Pm* is evaluated using the following formula:

$$P_m = \begin{cases} w \times \dfrac{N_{Elt}}{N_{ReqElm}} \ ; if \quad N_{Elt} < N_{ReqElm} \\ 1 \qquad\quad ; \ Otherwise \end{cases} \qquad (3)$$

Where N_{Elt} is the number of thematic, geospatial, or temporal elements of metadata which are semantically related and required by users, N_{ReqElm} is the total number of required metadata elements, and *w* a predefined value between 0 and 1 which indicates the weight of

each type (i.e., thematic, geospatial, and temporal) for end-users. If N_{Elt} is equal or larger than N_{ReqElm}, the metadata are perfectly relevant and has 1 as value.

2- Freshness of metadata. This indicator indicates the degree of freshness of the metadata of the geospatial datacubes. It is evaluated according to the age of metadata with regard to their lifetime. The age of metadata is the time passed since the date of metadata definition (*Tdef*) until the desired date of freshness of this metadata (*Treq*). The *lifetime* is the number of years after which the metadata will not be valid anymore. Freshness of metadata *Am* is evaluated as follow:

$$A_m = \begin{cases} 1 - \dfrac{\left|T_{req} - T_{def}\right|}{DV} \; ; \; if \quad \left|T_{req} - T_{def}\right| < DV \\ 0 \quad ; \qquad Otherwise \end{cases} \tag{4}$$

Where *DV* is the *liftime* of the metadata. A low value of the freshness decreases the fitness-for-use of metadata. This value decreases when its age increases. The *lifetime* and the date of metadata definition can be provided by the metadata producer.

The way of presenting these indicators to the user has a great importance on decision-making. Generally, decision-support systems use a restricted number of indicators (Few, 2006). Accordingly, we suggest to present only two indicators to the user: one for schema quality and the other for metadata quality. If the user would like to get more information about the quality of schema or the quality of metadata, the user would go into each of the previously defined indicator. The two indicators are calculated as follows:

$$Q_s = \frac{\sum (a \times I_s)}{n} \tag{5}$$

$$Q_m = \frac{\sum (b \times I_m)}{m} \tag{6}$$

Where *a* and *b* are predefined values between 0 and 1 that indicate the weight of to each quality indicator of schema and metadata respectively. *Is* and *Im* refer, respectively, to the

value of each schema and metadata indicator. The variables *n* and *m* are the numbers of the indicators.

In the following section, we propose a general framework to support agents in responding to the risk of data misinterpretation. We will show an example of how the proposed indicators can help making appropriate decisions about the risks of data misinterpretation.

Responding to the risks of data misinterpretation

This section proposes an approach to help the end-user to make appropriate decisions about the risk of data misinterpretation in the context of the interoperability between geospatial datacubes. The approach consists of proposing a general framework that presents the previously defined indicators to agents in an intuitive way in order to help them making appropriate decisions about the risks of data misinterpretation. The framework is based on the well established possibilities of responding to the risks (i.e., reducing, absorbing, or transferring the risks).

General framework to respond to the risks of data misinterpretation

The framework consists of five successive phases of analysing and responding to the risks of data misinterpretation. These phases correspond to the proposed categorization of semantic heterogeneity, from the more general level to the most detailed level: cube-to-cube, measure-to-measure, dimension-to-dimension, hierarchy-to-hierarchy, and level-to-level. At each level, the intervener analyzes the fitness-for-use of metadata and of schema based on the previously identified indicators, and makes one of the following decisions (see Figure 5):

- To suspend the interoperability process of the geospatial datacubes if it presents a high risk leading to harmful consequences. In this case, the intervener is invited not to continue with the remaining levels.

- To continue the interoperability process of the geospatial datacubes if there is no risk of data misinterpretation or if the risks are low. Consequently, two decisions can be made:

- Solving the causes of the risks.
- Doing nothing to solve the causes of the risks, and enduring the risks if they do not significantly affect data use.

The proposed indicators play a key role within the proposed framework. They allow to identify the risks of data misinterpretation and to draw some conclusions about them. More specifically, these indicators have three principal aims:

1. First, to help the intervener to identify the risks of data misinterpretation. In fact, at each level of the proposed framework, while a good quality of metadata and of schema indicates it is less likely to have a risk of data misinterpretation, a poor quality indicates a higher risk.
2. Second, to help agents to make appropriate decisions at each level (to suspend the interoperability process, to solve or endure the problems). For example, if the two heterogeneous elements of different models, which are essential for the interoperability, have a poor quality, the user can be advised not to consider both elements in any interoperability result.
3. Third, to help the intervener to solve the problems at each level of the proposed framework (cube-to-cube, measure-to-measure, dimension-to-dimension, hierarchy-to-hierarchy, level-to-level).

At the end of each level, according to the decision made, the intervener should write a report explaining the followings: the reasons of the suspension, how the problem was solved and the noted comments, or the reasons for which the agent decided to endure the problems.

The risk identification and management are carried out according to a hierarchical top-down approach which has two advantages:

1. The approach allows the intervener to make relevant decisions about the risks of data misinterpretation at an early stage of the interoperability process allowing them to put less time and effort dealing with such risks. At

the end of each level, interveners can suspend the interoperability before going into details and hence reduce the costs of the interoperability. Interveners can also continue the interoperability by taking into account the observations made at the general level to better deal with the detailed levels.

2. The proposed approach is in accordance with the mental model of human (Yougworth, 1995). Indeed, this framework is based on a hierarchical structure which is one of the essential principles of human cognition. This principle stipulates that humans gather data in categories according to their own knowledge (Mennis et al., 2000). These categories are organized in a hierarchical way in order to allow the maximum re-use of data with the minimum effort (Rosch, 1978).

Figure 5 illustrates the proposed framework. Based on the previously identified indicators of the fitness-for-use of metadata and schema, the agent starts by identifying the risks of data misinterpretation at the more general level (i.e., the cube level) and making decisions to solve their causes (e.g., problems of semantic heterogeneity).

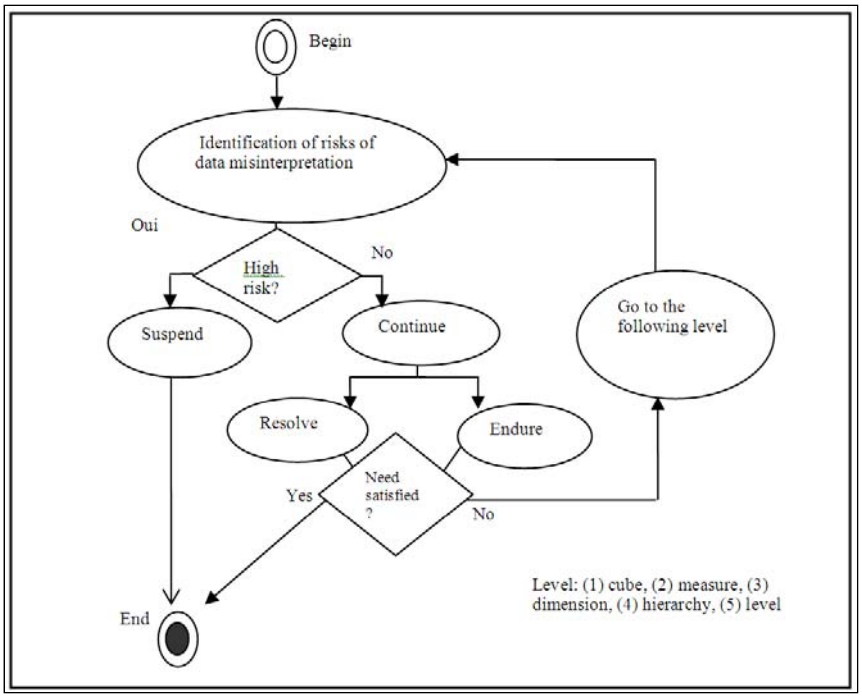

Figure. 5. A general framework to respond to the risks of data misinterpretation.

We should note that the problems of heterogeneity at the measure level must be dealt with before investigating the problems at the dimension level. The reason is that measures are the subject of analysis of the datacubes while dimensions represent the context of this analysis. After analyzing the risks at the dimension level, the user identifies the risks at the hierarchy level and makes decisions to solve them. Later on, we provide a continuation of the example presented in Figure 2 to show how the proposed framework can help to deal with the risks of data misinterpretation. .

Along with the content of the resulting message at each level, in the next section, we define a set of symbols indicating the degree of the risk of data misinterpretation during the interoperability process. This allows interveners to intuitively understand the message.

Symbolic notation of quality indicators

In order to help interveners to intuitively understand the degree of such a risk, we provide a set of warnings which are based on standard danger symbols proposed by ANSI Z535.4

(2006) and previously used in the geospatial domain by Levesque et al. (2007). These symbols are enriched by other symbols to suitably assist interveners in geospatial datacubes interoperability. The symbols show the degree of the risk of data misinterpretation at each level, and thereby stimulate appropriate responses to such a risk. For example, if the warning is 'Danger', it would be better to suspend the interoperability process which may lead to a considerable risk. Table 2 shows an example of how warnings can be predefined with regards to quantitative values of the fitness-for-use of geospatial datacube model.

Quality value	Representation of the quality indicators	Significance of the symbol
Q = 0		Interveners are advised to stop the process of interoperability.
0 < Q < = 0.25		There is a high risk if the intervener decides to continue integration.
0.25 < Q < = 0.5		The user is informed about the existence of potential risks.
0.5 < Q < = 0.75		Interveners are invited to pay attention if they decide to continue the interoperability.
0.75 < Q < 1		Information will be shown to interveners.
Q = 1		Interveners are invited to continue to a more detailed level.

Table. 2. A definition of the symbols according to the quality value.

To the above list, we add: the symbol  which invites the user to continue the evaluation of others semantically related elements, and the symbol  which indicates an absence of information about the semantically related elements (e.g., the absence of metadata), and thus the impossibility of evaluating their fitness-for-use.

In the following section, we provide a continuation of our example to show how the proposed approach can be used to deal with such risks.

Example of application

We continue with our example, i.e., two geospatial datacubes presented in Figure 2 and the requirements of the end-user in the form of a geospatial datacube model illustrated in Figure 4. In this example, we suppose that the goal of the interoperability is to define a common model that helps end-users to navigate simultaneously through different geospatial datacubes.

We evaluate the indicators at each level (cube-to-cube, measure-to-measure, dimension-to-dimension, hierarchy-to-hierarchy, level-to-level). To simplify the evaluation, we set the value 1 for *a* and *b* (which indicate the weight of each quality indicator of schema and metadata respectively; see equations (6) and (7)), and that for all levels.

For the two first levels (cube-to-cube and measure-to-measure), there is no problem of heterogeneity. Therefore, the symbol ⬅ will be shown for both levels.

Dimension-to-dimension:

There is a semantic heterogeneity between two dimensions *Administrative Region* of datacube C_1 and *Region* of datacube C_2, therefore the quality of both dimensions is evaluated for their structure and their metadata:

a- Schema heterogeneity:

- *Relevance of the number of hierarchies*

The dimension *Administrative Region* of the datacube C_1 has only one hierarchy: (*Country*, *Province*, *Territory* and *City*), but the dimension *Region* of the datacube C_2 has two hierarchies: (*Country*, *Province*, *Territory*, and *City*) and (*Country*, *State*, and *City*). Since the user needs only one hierarchy: (*Division Administrative*), the quality indicator *Relevance of the number of hierarchies* is evaluated according to the formula (1):

For C_1, Ps (*Administrative Region*) = 1/1 = 1

For C_2, since the number of hierarchy of the dimension *Region* is larger than that of the dimension semantically related of the required datacube C_3 (*Administrative Division*), then:

P_S *(Region)* = 1

According to the formula (5) the quality of datacubes schemas:

Q_S *(Administrative Region)* = 1

Q_S *(Region)* = 1

Consequently, the symbol ⬅ will be shown for the two dimensions *Administrative Region* and *Region*.

b- Metadata heterogeneity:

- *Relevance of metadata*

Metadata associated with the dimension *Administrative Region* of C_1 have two elements which are semantically related to the elements required by the user (the spatial coverage and the year of creation of the dimension). On the other hand, metadata associated with the dimension *Region* of C_2 have only one element which is semantically related to a required element (the spatial coverage of the dimension). If we consider that the weight of spatial information is 1, then based on formula (3):

P_m *(Administrative Region)* = 2/2 = 1

P_m *(Region)* = 1/2 = 0.5

- *Freshness of metadata*

Metadata of the dimensions *Administrative Region* of C_1 and *Region* of C_2 were created respectively in 2002 and 1982. Moreover, these two dimensions have the same lifetime: 30 years. Consequently, according to formula (4):

A_m (*Administrative Region*) = 1 - (2005-2002/30) = 0.9

A_m (*Region*) = 1 - (2005-1982/30) = 0.23

According to the formula (6) the quality of metadata:

Q_m *(Administrative Region)* = (1 + 0.9) /2 = 0.95

Q_m *(Region)* = (0.5 + 0.23) /2 = 0.36

Consequently, the symbol will be shown for the dimension *Administrative Region*. On the other hand, the symbol WARNING will be shown for the dimension *Region*. The user is then invited to be careful when considering the dimension *Region* in the process of interoperability, and to evaluate the quality of the detailed levels of these two dimensions (i.e., hierarchy) by taking into account the difference in quality of these dimensions.

Hierarchy-to-Hierarchy:

There is a semantic heterogeneity between the hierarchies (H_1: *City*, *Province*, *Territory* and *Country*) of dimension *Administrative Region* of datacube C_1 and (H_2: *City*, *Province*, *Territory* and *Country*) of the dimension *Region* of the datacube C_2, then we evaluate the quality of schema and metadata of the two hierarchies with regards to the requirement's hierarchy (H_3: *City*, *Province*, *Territory* and *Country*).

a- Schema heterogeneity:

- *Relevance number of levels*

Each hierarchy (H_1, H_2 and H_3) contains 4 levels. According to the formula (1):

P_S *(*H_1*,* H_3*)* = 4/4 =1

P_S *(*H_1*,* H_3*)* = 4/4 = 1.

- *The order of levels*

The hierarchy of the dimension *Administrative Region* has the following order: (*City* < *Province*), (*City* < *Territory*), (*Province* < *Country*) and (*Territory* < *Country*).

The hierarchy of dimension *Region* has the following order: (*City* < *Province*), (*City* < *Territory*), (*Province* < *Country*) and (*Territory* < *Country*). The hierarchy of dimension *Administrative Division* of the datacube C_3 has the following order: (*City* < *Province*) and (*Province* < *Country*).

For the datacube C_1, according to expression (2):

The levels *City*, *Province* (or *Territory*) and *Country* (C_1) are respectively, semantically related to the levels *City*, *Province* and *Country* (C_3). Moreover, the level orders *City* < *Province* (in C_1) and *City* < *Province* (in C_3), therefore elementary relevance of the order (*City* < *Province*) o_s = 1. Also, o_s (*City* < *Territory*) = o_S (*Province* < *Country*) = o_S (*Territory* < *Country*) = 1. Therefore, relevance of the structure is as follows:

$$O_S\,(H_1, H_3) = (1+1+1+1)\,/4 = 1$$

Similarly, the relevance of the hierarchy (*City*, *Province*, *Territory* and *Country*) of the datacube C_2 is calculated based on the expression (2):

$$O_S\,(H_2, H_3) = (1+1+1+1)\,/4 = 1$$

Finally, according to the formula (5):

$$Q_S\,(H_1, H_3) = 1$$

$$Q_S\,(H_2, H_3) = 1$$

Thus, the quality of schema of both hierarchies is very good. Consequently, the symbol  will be shown and the intervener is invited to continue with the remaining level.

b- Metadata heterogeneity:

- *Relevance of the metadata*

Metadata associated with the hierarchy (*City*, *Province*, *Territory* and *Country*) of the datacube C_1 contain an element which is semantically related to the required element (i.e., using the "multiple representation"). On the other hand, the metadata associated with the hierarchy (*City*, *Province*, *Territory* and *Country*) of C_2 do not contain any element that is semantically related to the required element. If we consider that the weight of spatial information is = 1, then, according to the formula (3):

$$P_m\,(H_1, H_3) = 1/1 = 1$$

$$P_m\,(H_2, H_3) = 0/1 = 0$$

Also, according to the formula (6):

$$Q_m\,(H_1, H_3) = 1/1 = 1$$

$$Q_m\,(H_2, H_3) = 0/1 = 0$$

Therefore, the symbol  for the hierarchy (*City*, *Province*, *Territory* and *Country*) of C_1 and the symbol 

for the hierarchy (*City*, *Province*, *Territory*, and *Country*) of C_2 will be shown. Accordingly, selecting the second hierarchy creates risks to harm the interoperability between geospatial datacubes. Thus, the user is invited to continue to evaluate only the levels of the first hierarchy (*City*, *Province*, *Territory* and *Country*) of C_1.

Heterogeneity Level-to-Level:

a- Schema heterogeneity (Level *City* of C_1)*:*

- *Relevance of the number of attributes*

The number of attributes of the level *City* of the datacube C_1 is one (*City-name*). Since the user needs two attributes for this level (*City-name* and *Surface*), then according to the formula (1):

$$P_S\,(City, City) = 1/2 = 0.5$$

- *Relevance of the geometric primitive*

In C_1, each member of the level *City* is represented by a point (0D), whereas the users need a polygon (2D) to represent the members of the same level *City*. According to the Table 1:

$$P_P\,(City, City) = 0.5$$

Then, according to the formula (5), the structure quality of the level *City* of C_1 is calculated as follow:

$$Q_S\,(City, City) = (0.5+0.5)\,/2 = 0.5$$

Consequently, the quality is reasonably satisfactory, hence the symbol 

is shown to inform the user about potential risks that may occur when considering this level. Based on the proposed framework, the user can decide to solve the problems related to this level or to endure the potential consequences of these problems.

b- Metadata heterogeneity (Level City of C_1):

- *Relevance of the metadata*

Metadata of the datacube C_1 contain two elements (the spatial referencing system and the scale of representation) which are semantically related to the required elements. If we consider that the weight of spatial information is 1, then according to the formula (3):

$$P_m\,(City, City) = 2/2 = 1$$

According to the formula (6), the quality of the metadata of the level *City* of C_1 is calculated as:

$$Q_m\,(City, City) = (1+1)\,/2 = 1$$

Consequently, the structure quality of the level *City* of C_1 is quite satisfactory. The symbol will be illustrated.

Similarly, the qualities of the levels *Province*, *Territory* and *Country* datacube C_1 are evaluated based on the formulas (5) and (6):

$$Q_S(Province, Province) = 1$$

$$Q_m(Country, Country) = 1$$

We should notice that, if an element of one of the datacubes sources is not semantically related to any other element of another datacube, and that it fits the user's requirement, then this element (measure, dimension, hierarchy, level) is integrated in the common model. The dimensions *Age category* and *Forest* in our example are of this case.

Figure 6 shows an example of model which could be obtained to enable the interoperability between geospatial datacubes C_1 and C_2 according to the different levels of the general framework and using the proposed quality indicators.

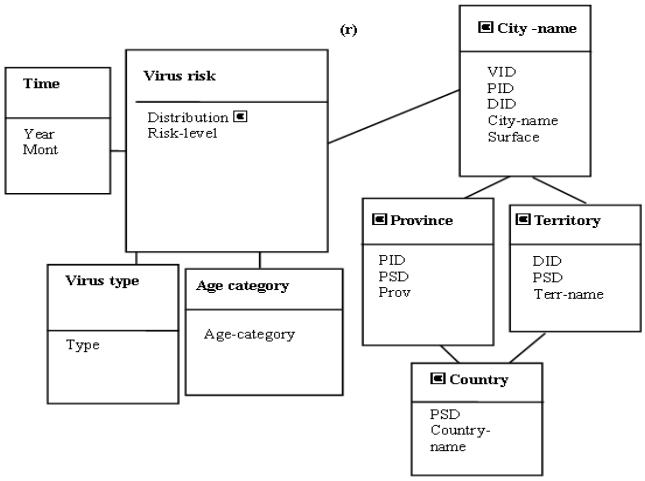

Figure 6. Example of the resulting common model.

Usefulness of the interoperability of geospatial datacubes for environmental applications

In many situations, such as a simultaneous and rapid intervention in environmental emergencies using different geospatial datacubes, a rapid acquisition of decisional data for

environmental applications, and an interactive analysis of environmental factor changes, we may need to interoperate several heterogeneous datacubes to support environmental applications. In the following, we provide some typical examples of these situations.

1. *Simultaneous and rapid intervention in environmental emergencies using different geospatial datacubes*: Users intervening in environmental emergency situations may need to access and navigate simultaneously through heterogeneous geospatial datacubes. Navigating separately through each datacube would be an arduous work for users, since they likely need to make extra efforts to manually resolve the problems of heterogeneity between datacubes (e.g., comparing the meaning of concepts and establishing a mapping between them). The principal aim of interoperability is to automatically overcome such differences which can considerably facilitate the navigation task and, hence, facilitate human intervention in emergency situation. For example, a health organization willing to analyze the risk of the West Nile virus to the population, may need to simultaneously navigate two geospatial datacubes; one containing data related to water bodies and the other containing data related to the location of dead birds reported by the population. The interoperability between these geospatial datacubes would enable, for example, to create a common multidimensional structure that allows to rapidly navigate through data stored in both geospatial datacubes. Our proposed approach can help intervener in the interoperability process to define such common structure.

2. *Rapid acquisition of decisional data for environmental applications*: While data in datacubes are usually collected from legacy systems, they can be imported from other heterogeneous datacubes (Bédard et al. 2001). We may need to rapidly acquire data from different heterogeneous geospatial datacubes to be used for environmental applications. For example, we may need to rapidly insert data in a geospatial datacube which contains data about forest stand from another geospatial datacubes that contains soil type of the same geographic region. The proposed approach in this paper can make intervener aware of potential

heterogeneity problems that may occur during data insertion, and can help him/her to make appropriate decisions to overcome such problems.

3. *An interactive analysis of environmental factor changes*: In order to analyze the changes of environmental factors (e.g., air pollution evolution), we may need to compare data describing these factors at different epochs, and stored in geospatial datacubes build also at different epochs. Interoperating such datacubes would enable to interactively compare data and analyze these factors' changes. For example, we may need to rapidly compare air pollution changes following a volcanic eruption. For that, we interoperate two existing geospatial datacubes for analyzing air pollution one based on data before the volcanic eruption and another one using data after eruption. Based on this study, we can realize the impact of the volcanic eruption on the air pollution. In our approach, we proposed a set of quality indicators that can help intervener in the interoperability process to compare the content of both geospatial datacubes.

Our proposed approach can be used in any of the above mentioned or similar situations to solve the problems arisen during the process of interoperability between geospatial datacubes. As the above examples illustrate, many of these situations can happen in environmental applications.

Conclusions

The interoperability between geospatial datacubes faces risks of data misinterpretation that may hinder the aim of strategic decision-making. These risks are typically caused by the semantic heterogeneity of the content of geospatial datacubes. In this paper, we proposed an approach to deal with such risks. We introduced a set of indicators of the fitness-for-use of conceptual model of datacube (i.e., quality of schema and of metadata with regard to a specific use), and quantitatively evaluated them. These indicators facilitate the identification and the evaluation of the risks, and hence, making users aware of the risks and their severity. In fact, while a good fitness-for-use of metadata and of schema indicates it is less likely to faulty interpreting data or being uncertain about its interpretation, a poor quality indicates a higher risk.

In addition, we proposed a general framework that can be used by interveners to make appropriate decisions about the risks that may occur during the interoperability process of geospatial datacubes. The framework is based on a hierarchical top-down that corresponds to the proposed categorization of semantic heterogeneity (cube-to-cube, measure-to-measure, dimension-to-dimension, hierarchy-to-hierarchy, and level-to-level). We also demonstrated the usefulness of the proposed approach in environmental applications.

We should remind that the set of the proposed indicators and the proposed framework do not aim being exhaustive or precise but rather at helping the intervener to make appropriate decisions to enhance the interoperability between geospatial datacubes. Such a method is frequently used in several fields like epidemiology and ecology which involve factors that are difficult, even impossible, to evaluate in an exhaustive and precise way.

The example of application showed how the framework as well as the proposed indicators can help interveners to make appropriate decisions about the suspension or the continuation of interoperability process. The framework and the indicators constitute a base for future works dealing with the interoperability between geospatial datacubes, but also with interoperability involving other information systems.

Reference

Agumya, A., & Hunter, G.J. (2002). Responding to the consequences of uncertainty in geographical data. *IJGIS,16(5), 405-417*.

ANSI Z535.6 (2006). *American national standard for product safety signs and labels*.

Bédard, Y. (1986). A Study of data using a communication based conceptual framework of land information systems. *Le Géomètre Canadien, 40(4), 449-460*.

Bédard, Y., Merrett, T. & Han, J. (2001). Fundamentals of spatial data warehousing for geographic knowledge discovery. In H.J. Miller & J. Han (Ed.), *Geographic data mining and knowledge discovery*.

Bédard, Y., Rivest, S., & Proulx, M.J. (2005). Spatial on-line analytical processing (SOLAP): concepts, architectures and solutions from a geomatics engineering perspective. In W.R. Koncillia (Ed.), *Data warehouses and OLAP: Concepts, architectures and solutions*.

Bédard, Y., & Larrivée, S. (2008). Spatial database modeling with pictrogrammic languages. In S. Shekhar & H. Xiong (Ed.), *Encyclopedia of GIS* (pp. 716-725). Springer-Verlag.

Bishr, Y. (1998). Overcoming the semantic and other barriers to GIS interoperability. *IJGIS 12, 299–314*.

Boin, A.T. (2008). *Exposing Uncertainty: Communicating spatial data quality via the Internet*. Ph.D. Dissertation. University of Melbourne, pp. 197.

Brodeur, J., *Interopérabilité des données géospatiales: élaboration du concept de proximité géosémantique*. Ph.D. Dissertation. Université Laval, pp. 247.

Cherfi, S.S., Akoka, J., & Comyn-Wattiau, I. (2007). Perceived vs. measured quality of conceptual schemas: an experimental comparison. *ACM Int. C.Proceeding Series, 334, 185-190*.

Devillers, R., Bédard, Y., Jeansoulin, R., & Moulin, B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *IJGIS, 21(3):261-282*.

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media.

Frank, A.U. (2007). Data quality ontology: An ontology for imperfect knowledge. *COSIT 2007, LNCS 4736, 406–420*.

Genero, M., Manso, E., Visaggio, A., Canfora G., & Piattini, M. (2007). Building measure-based prediction models for UML class diagram maintainability. *Empir Software Eng.,12, 517–549.*

Kuhn, W. (2005). Geospatial semantics: Why, of what, and how. *Journal on Data Semantics III, 1-24*.

Levesque, M.-A., Bédard, Y., Gervais, M., & Devillers, R. (2007). Towards managing the risks of data misuse for geospatial datacubes de données. *ISSDQ 2007*.

Mennis, J.L., Peuquet, D. J., & Qian, L. (2000). A conceptual framework for incorporating cognitive principles into geographical database representation. *IJGIS, 14(6),501-520.*

Moody, D.L. (2005). Theoretical and practical issues in evaluating the quality of conceptual models: Current state and future directions. *Data & Knowledge Engineering, 15(3), 243-276*.

Rafanelli, M. (2003). *Multidimensional databases: Problems and solutions*, Idea Group Publishing.

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

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Ed.), *Cognition and Categorization* (pp. 27-77).

Sboui, T., Bédard, Y., Brodeur, J., & Badard, T. (2007). A conceptual framework to support semantic interoperability of geospatial datacubes. *Proceedings of ER/2007 SeCoGIS workshop, LNCS 4802* (pp. 378-387), Springer.

Staub, P., Gnagi, H.R., & Morf, A. (2008). Semantic interoperability through the definition of conceptual model transformations. *Transactions in GIS, 12(2),193-207*.

Yougworth, P. (1995). OLAP spells success for users and developers. *Data Based Advisor, 38-49*.