

TAREK SBOUI

**A CONCEPTUAL FRAMEWORK AND A RISK
MANAGEMENT APPROACH FOR
INTEROPERABILITY BETWEEN GEOSPATIAL
DATACUBES**

Thèse présentée
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de doctorat en Sciences géomatiques
pour l'obtention du grade de Philosophiae Doctor (Ph.D.)

DEPARTEMENT DES SCIENCES GEOMATIQUES
FACULTÉ DE FORESTERIE ET DE GÉOMATIQUE
UNIVERSITÉ LAVAL
QUÉBEC

2010

Résumé

De nos jours, nous observons un intérêt grandissant pour les bases de données géospatiales multidimensionnelles. Ces bases de données sont développées pour faciliter la prise de décisions stratégiques des organisations, et plus spécifiquement lorsqu'il s'agit de données de différentes époques et de différents niveaux de granularité. Cependant, les utilisateurs peuvent avoir besoin d'utiliser plusieurs bases de données géospatiales multidimensionnelles. Ces bases de données peuvent être sémantiquement hétérogènes et caractérisées par différents degrés de pertinence par rapport au contexte d'utilisation. Résoudre les problèmes sémantiques liés à l'hétérogénéité et à la différence de pertinence d'une manière transparente aux utilisateurs a été l'objectif principal de l'interopérabilité au cours des quinze dernières années. Dans ce contexte, différentes solutions ont été proposées pour traiter l'interopérabilité. Cependant, ces solutions ont adopté une approche non systématique. De plus, aucune solution pour résoudre des problèmes sémantiques spécifiques liés à l'interopérabilité entre les bases de données géospatiales multidimensionnelles n'a été trouvée.

Dans cette thèse, nous supposons qu'il est possible de définir une approche qui traite ces problèmes sémantiques pour assurer l'interopérabilité entre les bases de données géospatiales multidimensionnelles. Ainsi, nous définissons tout d'abord l'interopérabilité entre ces bases de données. Ensuite, nous définissons et classifions les problèmes d'hétérogénéité sémantique qui peuvent se produire au cours d'une telle interopérabilité de différentes bases de données géospatiales multidimensionnelles. Afin de résoudre ces problèmes d'hétérogénéité sémantique, nous proposons un cadre conceptuel qui se base sur la communication humaine. Dans ce cadre, une communication s'établit entre deux agents système représentant les bases de données géospatiales multidimensionnelles impliquées dans un processus d'interopérabilité. Cette communication vise à échanger de l'information sur le contenu de ces bases.

Ensuite, dans l'intention d'aider les agents à prendre des décisions appropriées au cours du processus d'interopérabilité, nous évaluons un ensemble d'indicateurs de la qualité externe (fitness-for-use) des schémas et du contexte de production (ex., les métadonnées). Finalement, nous mettons en œuvre l'approche afin de montrer sa faisabilité.

Abstract

Today, we observe wide use of geospatial databases that are implemented in many forms (e.g., transactional centralized systems, distributed databases, multidimensional datacubes). Among those possibilities, the multidimensional datacube is more appropriate to support interactive analysis and to guide the organization's strategic decisions, especially when different epochs and levels of information granularity are involved. However, one may need to use several geospatial multidimensional datacubes which may be semantically heterogeneous and having different degrees of appropriateness to the context of use. Overcoming the semantic problems related to the semantic heterogeneity and to the difference in the appropriateness to the context of use in a manner that is transparent to users has been the principal aim of interoperability for the last fifteen years. However, in spite of successful initiatives, today's solutions have evolved in a non systematic way. Moreover, no solution has been found to address specific semantic problems related to interoperability between geospatial datacubes.

In this thesis, we suppose that it is possible to define an approach that addresses these semantic problems to support interoperability between geospatial datacubes. For that, we first describe interoperability between geospatial datacubes. Then, we define and categorize the semantic heterogeneity problems that may occur during the interoperability process of different geospatial datacubes. In order to resolve semantic heterogeneity between geospatial datacubes, we propose a conceptual framework that is essentially based on human communication. In this framework, software agents representing geospatial datacubes involved in the interoperability process communicate together. Such communication aims at exchanging information about the content of geospatial datacubes.

Then, in order to help agents to make appropriate decisions during the interoperability process, we evaluate a set of indicators of the external quality (fitness-for-use) of geospatial datacube schemas and of production context (e.g., metadata). Finally, we implement the proposed approach to show its feasibility.

Acknowledgments

First and foremost I thank God for giving me the stamina, ability and knowledge to complete my Ph.D. thesis.

During this thesis, I was supported by many people whom I wish to thank for their sincere contribution. First of all, I would like to express my deepest appreciation and gratitude to my thesis supervisor Dr Yvan Bédard. Yvan was abundantly helpful and critical advisor, who gave me the freedom needed to develop my research. I learned a lot from him both scientifically and personally. Yvan has been available to help me despite his many responsibilities as the - NSERC Industrial Research Chair of Geospatial Database for Decision Support (2004 - 2010).

I also would like express my sincere gratitude to Dr Jean Brodeur and Dr Thierry Badard. Dr Jean Brodeur has provided a great deal of encouragement and support during my Ph.D. studies. In addition, he encouraged me to get involved in two projects at the Centre for Topographic Information – Sherbrooke where I gained insight into the interoperability paradigm. Thierry Badard has been a source of good advice and encouragement. His ideas and suggestions have greatly helped me in my thesis research. I would like to thank both of them for their friendship, supervision and invaluable assistance.

Special thanks to Dr Jacynthe Pouliot for accepting to perform the “pré-lecture” of my thesis and being a member of the Jury. Jacynthe’s valuable feedback comments helped me to improve the quality of my thesis. I also would like to express my sincere gratitude to Dr François Pinet for accepting to be a member of the Jury.

Also, I would like to thank the NSERC Industrial Research Chair of Geospatial Database for Decision Support for its support. The chair was financed by the Natural Sciences and Engineering Research Council of Canada, Laval University, Hydro-Québec, Re-search and Development Defence Canada, Natural Resources Canada, Ministère des Transports du Québec, KHEOPS Technologies, Intélec Géomatique, Syntell, Holonics, and DVP-GS.

In addition, I would like to thank all members of the Yvan Bédard’s research group. I feel fortunate for having being part of a friendly environment that made my Ph.D. program an

enjoyable and unforgettable experience. I particularly thank Eveline Bernier, Sonia Rivest, Suzie Larrivée, Mamane Nouri Sabo, Frédéric Hubert, John-William Cely-Pulido and Eve Grenier. Special thanks to my colleagues and friends Lotfi Bejaoui and Mehrdad Salehi, with whom I enjoyed working and talking.

Finally, I wish to express my love and appreciation to my father, brothers, and sisters for sharing the good and bad moments of my life, especially through the duration of my PhD studies.

*To the memory of my mother and my brother
To my father, sisters and brothers
To my friends*

Table of Contents

Résumé.....	i
Abstract.....	ii
Acknowledgments	iii
Table of Contents.....	vi
List of Tables	ix
List of Figures.....	x
Chapter 1: Introduction.....	2
1.1 Research context.....	2
1.2 Problem statement.....	6
1.3 Hypotheses of the research and objectives	15
1.3.1 Hypotheses.....	15
1.3.2 Objectives	15
1.4 Methodology.....	17
1.5 Structure of the thesis	25
Chapter 2: Literature Review.....	26
2.1 Introduction.....	26
2.2 Datacubes: Characteristics and Functionalities	26
2.2.1 From Transactional Databases to Datacubes	26
2.2.2 Geospatial datacubes.....	34
2.3 Interoperability: review of key concepts and existing approaches	36
2.3.1 Definition of interoperability.....	36
2.3.2 Types and levels of interoperability.....	37
2.4 Semantic interoperability: an analogy with human communication	40
2.4.1. Semantics of geospatial data.....	40
2.4.2. Semantic heterogeneity: a barrier for interoperability	41
2.4.2.1. Semantic heterogeneity in traditional databases.....	41
2.4.2.2. Semantic heterogeneity in geospatial databases	42
2.4.2.3. Semantic Heterogeneity in traditional (non-spatial) datacubes	47
2.4.3. Key notions of semantic interoperability.....	49
2.4.3.1 Human communication and interoperability	49
2.4.3.2 Ontology	53
2.4.3.3 Context.....	56
2.4.3.4 Metadata.....	58
2.4.3.5 Standards.....	59
2.4.3.6 Geospatial data quality.....	60
2.4.3.7 Semantic similarity	63
2.4.3.8 Perspective of the basic notions.....	64
2.5 Semantic interoperability – review of related works	64
2.5.1 Semantic interoperability of traditional databases.....	65
2.5.2 Semantic interoperability of geospatial databases	67
2.5.3 Semantic interoperability of datacubes.....	71
2.5.4 Perspective of existing approaches of interoperability	73
2.6 Conclusions.....	75
Chapter 3: Defining semantic interoperability between geospatial datacubes	77

3.1 Introduction.....	77
3.2 Need for interoperating geospatial datacubes.....	78
3.3 A new definition of interoperability between geospatial datacubes.....	81
3.4 Semantic heterogeneity in geospatial datacubes.....	88
3.5 Conclusions.....	99
Chapter 4: A conceptual framework for semantic interoperability between geospatial datacubes and associated risks.....	100
4.1 Introduction.....	100
4.2. A conceptual framework for the semantic interoperability between geospatial datacubes.....	100
4.3 Resolving semantic heterogeneity in geospatial datacubes.....	109
4.3.1 Using the GsP notion.....	109
4.3.2 MGsP: Extending the GsP notion to stress the semantics of geospatial datacubes elements.....	111
4.4 Limits of the ontology-based interpretation of geospatial concepts.....	118
4.5 Human intervention to support semantic interoperability between geospatial datacubes.....	121
4.6 Conclusions.....	123
Chapter 5: Fine-tuning the semantic interoperability between geospatial datacubes with a risk management approach based on the fitness-for-use of conceptual models.....	125
5.1 Introduction.....	125
5.2 Semantic interoperability between geospatial datacubes and risks of data misinterpretation.....	126
5.2.1 Risk and risk management.....	126
5.2.2 Overview of risks of data misinterpretation in geospatial datacubes interoperability.....	127
5.2.3 Causes of the risks of data misinterpretation.....	130
5.2.4. Managing the risks of data misinterpretation: an overview of approaches.....	131
5.2.4.1 Prevention approaches.....	131
5.2.4.2 Reactive approaches.....	131
5.3 An approach to identify and evaluate the risks of misinterpretation in semantic interoperability of geospatial datacubes.....	132
5.3.1 Fitness-for-use of geospatial datacube conceptual models to identify and evaluate the risks.....	133
5.3.1.1 The quality of schema.....	134
5.3.1.2 The quality of production context.....	135
5.3.2 Indicators to identify and evaluate the risks in the semantic interoperability between geospatial datacubes.....	136
5.3.2.1. Indicators for evaluating the external quality of geospatial datacube schema.....	138
5.3.2.2 Indicators for evaluating the fitness-for-use of production context.....	140
5.4 Responding to the risks of data misinterpretation in the semantic interoperability of geospatial datacubes.....	143
5.4.1 General framework to respond to the risks of data misinterpretation.....	145
5.4.2 An Algorithm to support decision-making in the semantic interoperability of geospatial datacubes.....	148
5.4.3 Symbolic notation of quality indicators.....	150

5.5. Example of application	151
5.6. Conclusions.....	155
Chapter 6: An experimentation of the proposed approach for the interoperability between geospatial datacubes	157
6.1. Introduction.....	157
6.2. Technology used	157
6.2.1 Software agents to mimic human communication.....	157
6.2.2 GsP tool to measure semantic similarity.....	159
6.3. MQIGDC implementation and experimentation	159
6.3.1 MQIGDC architecture	160
6.3.2 MQIGDC implementation	163
6.3.2.1 Environment of implementation	163
6.3.2.2 Interfaces of the MQIGDC prototype	163
6.3.3 Experimentation.....	169
6.3.3.1 Phases of experimentation	170
6.3.3.2 Summary of the results	178
6.4. Conclusions.....	181
Chapter 7: Conclusions and discussion	183
7.1 Summary.....	183
7.2 Contributions	183
7.3 Discussion.....	189
7.4 Final conclusions	191
7.5 Research perspectives	193
Bibliography	198
Annex A: An example of generic context elements	211
Annex B: A Conceptual Framework to Support Semantic Interoperability of Geospatial Datacubes	213
Annex C: Modeling the External Quality of Context to Fine-tune Context Reasoning in Geospatial Interoperability	218
Annex D: Example of application of managing the risks of data misinterpretation based on the fitness-for-use of conceptual models	227

List of Tables

Table 2.1: Transactional Databases versus Data Warehouses (Bédard and Han 2008).	27
Table 2.2: Levels of interoperability (adapted from (Goodchild et al. 1999)).	38
Table 5.1: Evaluating the relevance of the geometric primitive.	139
Table 5.2: A definition of the symbols according to the quality value.	150
Table 6.1: The fitness-for-use of datacubes elements.	178
Table 6.2: Supporting the GsP tool to fine-tune the interoperability process.	179
Table A.1: An example of generic context elements.	212

List of Figures

Figure 1.1: Same symbol, different perceptions.	10
Figure 1.2: Semantics of a polygon in different application.	19
Figure 1.3: UML activity diagram of the research methodology.	23
Figure 2.1: An example of a datacube and its elements.	29
Figure 2.2: The result of applying roll-up on the level <i>City</i>	30
Figure 2.3: The result of applying drill-down on the level <i>Province</i>	31
Figure 2.4: The result of slicing the dimension <i>Type of vehicles</i>	31
Figure 2.5: The result of applying dice on the member “Montreal”.	32
Figure 2.6 : The result of applying pivot on the datacube Figure 2.1.	33
Figure 2.7: A (a) star schema, (b) snowflake schema, and (c) fact constellation schema. ...	34
Figure 2.8: Semantics of a polygon in different application.	40
Figure 2.9: Polygon indicating different objects according to their graphic form (shape). ...	41
Figure 2.10: The geoid-ellipsoid-map transfer process (Bédard 2005).	44
Figure 2.11: Example of the spatial aggregation-generalization mismatch (Bédard et al. 2005).	46
Figure 2.12: Communication process (Schramm 1971).	51
Figure 2.13: communication between two agents A et B.	51
Figure 2.14: Schramm’s concepts.	51
Figure 2.15: Meaning triangle based on Ogden & Richards (Ogden and Richards 1923). ...	53
Figure 2.16: Data federation architecture (Sheth 1999).	66
Figure 2.17: Framework for integrating data warehouses (Bruckner et al. 2001).	71
Figure 3.1: Inserting a measure from another datacube.	83
Figure 3.2: Creating a new measure from two existing measures.	84
Figure 3.3: Choosing one of two existing measures.	84
Figure 3.4: Adding a dimension to datacube from another.	85
Figure 3.5: Creating a new dimension from two existing dimensions.	86
Figure 3.6: Choosing one of two existing dimensions.	86
Figure 3.7: Transformation of a simple hierarchy Territory for one country to a multiple hierarchy by adding the level <i>State</i>	87
Figure 3.8: Two models of datacubes (C1 and C2).	91
Figure 3.9: Temporal heterogeneity: fusion of municipalities <i>A</i> and <i>C</i> in 2006.	98
Figure 4.1: An overview of the communication between agents representing geospatial datacubes.	102
Figure 4.2: Conceptual framework between geospatial datacube agents.	104
Figure 4.3: The communication protocol for the proposed conceptual framework	108
Figure 4.4: 4-intersection multidimensional matrix as a specialization of the GsP.	113
Figure 4.5: 16 possible MGsP predicates.	115
Figure 4.6: Different interpretations in different contexts.	119
Figure 5.1: Risk of misinterpreting data (“The river 2 intersects with the river 1”).	128
Figure 5.2: Geospatial datacube model describing the needs of the end-user.	137
Figure 5.3: A general framework to respond to the risks of data misinterpretation.	147
Figure 6.1: General architecture for geospatial datacube interoperability.	160

Figure 6.2: Embedding fitness-for-use representation within the representation of geospatial concept in GsP.	162
Figure 6.3: The agent manager window.	164
Figure 6.4: The agent window.	164
Figure 6.5: The quality (fitness-for-use) window.	165
Figure 6.6: The fitness-for-use assessment window.	166
Figure 6.7: The indicator's value calculation interface.	166
Figure 6.8: An example of a suggestion message.	167
Figure 6.9 : Decision-based communication.	168
Figure 6.10: Suggestion to continue to the next level.	171
Figure 6.11: Evaluation of the fitness-for-use of the schema and metadata of the dimensions <i>Administrative region</i> and <i>Region</i>	172
Figure 6.12: Value of the relevance of the number of hierarchy of <i>Administrative region</i>	172
Figure 6.13: Value of the overall fitness-for-use of structure.	172
Figure 6.14: Value of the relevance of the number of hierarchy of <i>Administrative region</i>	173
Figure 6.15: Value of the freshness of metadata of to the dimension <i>Region</i>	173
Figure 6.16: Value of the overall fitness-for-use of metadata related to the dimension <i>Administrative region</i>	174
Figure 6.17: Making the stakeholder aware of the risk in considering the dimension <i>Region</i>	174
Figure 6.18: Context agent communicating with the dimension agents.	175
Figure 6.19 : Making the stakeholder aware of the risk of considering the hierarchy (<i>City</i> , <i>Province</i> , <i>Territory</i> , and <i>Country</i>) of <i>C2</i>	176
Figure 7.1: A propagation of the risks of data misinterpretation.	197

Chapter 1: Introduction

1.1 Research context

In the last two decades, we have witnessed important technological innovations (e.g., satellite technology and telecommunications systems). Moreover, there has been a significant increase in the number of information sources (different databases, Web sites, etc.). Decision support systems have been among the fields that have seen major advances (Turban and Aronson 2001, Bédard and Han 2008). Such systems are intended to provide support to strategic decision makers (analysts, executives, and managers) engaged in solving complex problems. Data warehouses are considered an efficient, integral part of modern decision support systems.

A data warehouse is a large repository of subject oriented, integrated, time varying, non-volatile collection of data (Chaudhuri and Dayal 1997). Data are typically extracted from different transactional sources, transformed to meet the business needs, and finally integrated and loaded into a unified data warehouse environment (Bédard and Han 2008). Data warehouses provide decision makers with aggregated and summarized historical data, which is generally derived from transactional data, and computed using a given function (e.g., count, sum, and avg) according to different levels of granularity. Such data provide significant insights to decision makers, as they can use it to analyze historical trends and exploit elements that affect their businesses. Data warehouses are often structured as datacubes, i.e., according to the multidimensional paradigm defined in the field of Business Intelligence (BI) (Gray et al. 1997). This paradigm enables making strategic decisions by supporting the user's mental model of data. It allows users to navigate aggregated and summarized data according to a set of dimensions with different hierarchies (Codd 1993). In addition, it allows strategic decision makers to gain insight into data by fast and interactive access to a wide variety of possible views of information using different tools such as the On-Line Analytical Processing (OLAP) tools (OLAP Council 1995). Some key

concepts of the datacube are dimensions, levels, members, measures, and facts (see chapter 2 for detailed definitions of these terms).

Furthermore, thanks to different data acquisition methods and technologies in geomatics such as GPS, remote sensing, photogrammetry, etc., large amounts of geospatial data are available. Geospatial data allows to visualize real world phenomena taking into account their geospatial characteristics (e.g., position, shape, and size), geospatial relationships (e.g., adjacency, connectivity, inclusion, proximity, and overlay), and geospatial distribution (e.g., concentrated, scattered, and grouped) (Chorley and Haggett 1967, Bilodeau 1991, Bédard et al. 2005). In addition, it is estimated that geospatial data constitute the major part of governmental information: according to Franklin (1992), 80% of governmental information has geospatial characteristics, a rate that has been accepted in the industry.

In order to combine the benefits of geospatial data and the efficiency of the datacubes in the decision making process, geospatial datacubes have been introduced (Han et al. 1998, Bédard and Han 2008, Shekhar et al. 2001, Rivest et al. 2005, Damiani and Spaccapietra 2006, Malinowski and Zimányi 2008). Geospatial datacubes integrate geospatial data and the multidimensional paradigm, and are recognized as one of the most promising components of decision-support systems (Rafanelli 2003, Bédard et al. 2005). In recent years, geospatial datacubes have attracted a lot of attention and several research topics in the realm of geospatial datacubes have been addressed (Papadias et al. 2002, Rivest et al. 2005, Choi and Luk 2008, Malinowski and Zimányi 2008, Salehi 2009). Examples of research topics include conceptual modeling, geospatial indexing, query processing, and exploitation of geospatial datacubes using tools such as spatial OLAP (SOLAP) tools.

It may happen that users (end-users (e.g., decision makers) or datacube developers) need to reuse several scattered geospatial datacubes at the same time to make strategic decisions or discover geospatial trends. 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.

Simultaneously using different geospatial datacubes encounters problems due to the fact that geospatial datacubes are most of the time heterogeneous. In fact, these datacubes, as it is the case for the other databases, are often developed using different techniques and their content are often described using different representations. Consequently, geospatial datacubes may differ in format and content (e.g., different schemas, different geospatial characteristics such as locations and geometries and levels of abstraction). The heterogeneity can be categorized into technical, structural, and semantic heterogeneity (Chatterjee and Segev 1992, Denk and Oropallo 2002). Technical heterogeneity involves differences in hardware, operating systems, and in database management systems or OLAP servers. Structural heterogeneity basically involves difference in database models (e.g., relational and object-oriented), differences in data types and data formats, and differences in measurement units (Denk and Oropallo 2002). Semantic heterogeneity, which is the focus of this research work, basically occurs when there is a difference in the meaning or interpretation of the same or related data (i.e., concepts having a similarity, generalization, or specialization relationship) (Sheth and Larson 1990, Denk and Oropallo 2002, Park and Ram 2004). Such difference is caused basically by the difference in data description. Data description refers the set of terms and predicates used to describe data. Typical examples of semantic heterogeneity are 1) to use the same data to represent different concepts (e.g., polygons may refer to a roof or to a ground of a residence on a map), and 2) to use different data to represent the same concept (e.g., polygons and points may refer to buildings on a map). Semantic heterogeneity may also involve differences of temporal validity and level of granularity (Denk and Oropallo 2002). Moreover, the lack or the inappropriateness of context information, which refers to the discordance in the amount of relevant context information associated with different datasets, may also be considered as contributing to semantic heterogeneity as it has direct impacts on the interpretation of the data. Compared to the two first categories (i.e., technical heterogeneity and structural heterogeneity), semantic heterogeneity, which is the main concern of this thesis, is the most difficult to overcome (Kashyap and Sheth 1996, Bishr 1998, Brodeur 2004).

Different ways can be used to deal with the problems related to the heterogeneity, and to simultaneously reuse heterogeneous geospatial datacubes, e.g., separately navigating through each geospatial datacube, and interoperating the datacubes to be reused. Navigating

separately through each datacube would be an arduous work for users, as they would try to overcome the heterogeneity problems that may exist between the datacubes. Overcoming such heterogeneity in a large-scale geospatial datacubes, if it is possible, requires an extensive effort (e.g., assigning meaning to data, and comparing similar representations) and different techniques (e.g., GIS tool to resolve geometric incompatibility of same objects). On the other hand, interoperability has been widely recognized as an efficient paradigm for joining different systems to facilitate an efficient exchange of information in many fields (information management, engineering technologies, etc.) (Brodie 1992, Bishr 1998, Harvey et al. 1999, Brodeur 2004, Staub et al. 2008), and can be used to facilitate an efficient reuse of heterogeneous geospatial datacubes. With the emergence of software agents¹, semantic interoperability has been viewed as the technical analogue to human communication (Brodeur 2004, Kuhn 2005). According to this view, each receiver agent tries to assign meaning to (i.e., interpret) the exchanged data as it has been originally intended by a source agent, and to adapt these data to his/her/its context.

In spite of successful initiatives of semantic interoperability of transactional databases and widespread use of standards, today's solutions do not address yet geospatial datacubes. Moreover, no existing approach or standard can deal with the multitude problems that may occur during the interoperability of information systems in general. As an example, we quote the GeoConnections project² in which we participated during the thesis's process. The goal of the project is to develop an interactive system that supports humans in searching for geospatial datasets over the Web. The system uses an ontology-based service to describe the content and localization of dataset's elements. However, two major problems were experienced: 1) generally the terms used by users and those employed by service provider are different (e.g., misspelled terms and synonyms). For instance, searching for cadastral data using the term "cadastre" as a keyword, the system found no dataset. However, we know that there are more than 30 datasets related to the cadastral subject. The system failure may be due to the fact that existing datasets are described in English or the system uses a different ontology than the user does. Also, 2) the systems

¹ A software agent is defined as "a component of a software and/or hardware which is capable of acting exactly in order to accomplish tasks on behalf of its user." (Nwana, 1996)

² [http://www.geomatics2009.com/en/component/hpjgestion/?task=preview&tmpl=component&cid\[0\]=1006](http://www.geomatics2009.com/en/component/hpjgestion/?task=preview&tmpl=component&cid[0]=1006)

may fail to interpret a given location, or to infer related locations. For instance, a search for some datasets available in Quebec does not return those available in Montreal.

In light of these difficulties, it seems relevant to study specific problems related to the interoperability between geospatial datacubes, and to investigate possibilities to better support such interoperability.

1.2 Problem statement

In order to deal with the heterogeneity problem, many approaches of interoperability between transactional geospatial databases have been developed (Brodie 1992, Goodchild et al. 1999, Bishr 1998, Harvey et al. 1999, Brodeur 2004, Staub et al. 2008) and many standards have been specified (OGC 2002, ISO/TC 211 2003). However, there is **no work on semantic interoperability between geospatial datacubes** that has been found in the literature. We should notice that for the interoperability of non-spatial datacubes, few approaches have been proposed (Bruckner et al. 2001, Mangisengi et al. 2001, Pedersen et al. 2002, Hümmer et al. 2003, Frank and Chen 2005). However, these approaches have been focusing on the structural and technical aspects rather than the semantic aspects of datacube content (details will be given in chapter 2).

Furthermore, the existing approaches for semantic interoperability of transactional databases, which will be analyzed in detail in chapter 2, focus on transactional aspects of data (i.e., normalised, detailed, application-oriented data), and do not intend to support users interested in aggregated/summarized data for their strategic decision-making. They do not focus on presenting the interoperability result to the users in a timely fashion and an appropriate data organisation that would allow an intuitive use of data, and hence may not be completely suitable for strategic decision-making needs. For example, the approach proposed by Lutz and Klien (2006) to discover geospatial data is based on semantic matchmaking between concepts representing geographic feature types (i.e., classes of geographic objects with common characteristics) on the one hand and concepts representing user's query on the other hand. The result returned to end-users is a set of matched concepts with their attributes organized in a simple table form without any

appropriate pre-organization for strategic decision-making needs like the one provided by the multidimensional paradigm.

We should note that obtaining an intuitive data view from existing approaches of interoperability requires end-users to either write complex SQL queries or use an SQL generator to convert the table outputs into a more intuitive format. For example, in the framework proposed by Nambiar et al. (2006), users may need to be well skilled in issuing SQL queries in order to search for ontology concepts.

On the other hand, while the interoperability between transactional databases and the interoperability between datacubes have a common *raison d'être* which is data reuse, the latter is oriented towards supporting users in their strategic decision-making process. That is, any interoperability process of geospatial datacubes would take into account the ability to provide a strategic decision support (i.e., provide users with relevant elements to make strategic decisions which typically rely on trends and summarized crosstab data). In other words, any result of such interoperability (e.g., federated datacube and common schema) should support the user's mental model of data and allow him/her to rapidly navigate aggregated and summarized data. For example, an interoperability of the above mentioned datacubes (*water bodies* and location of *dead birds*) should support a rapid and multidimensional access to historical data stored in both datacubes by providing, for instance, a common multidimensional schema that allows to access and rapidly navigate both datacubes as if they were only one.

Moreover, the elements of interest to be compared when interoperating datacubes are different from those of transactional databases. First, while the semantic interoperability of transactional databases deals with tables, attributes and values, semantic interoperability of datacubes deals with dimensions, hierarchies, levels, members, measures and facts. Second, unlike the semantic interoperability of transactional databases, the semantic interoperability of datacubes particularly stresses the importance of dealing with the semantics of aggregation and generalization relationships, the semantics of summarizing methods and algorithms, the semantics of summarizability conditions, the semantics of cross-tabulations for every level of details and every member of the datacube dimensions and the semantics of geospatial hyper-cells which describe a model for a number of facts. Third, the semantic

interoperability of datacubes deals with the context associated with decisional data. Such context may contain complex elements such as multidimensional integrity constraints which are assertions typically defined in order to prevent the insertion of incorrect data into datacubes³ (Salehi 2009). A detailed comparison of interoperability between transactional databases with the interoperability between datacubes will be provided in chapter 3.

Furthermore, since data have no built-in intrinsic meaning, interpreting the exchanged data as it has been originally intended and to adapt it to the current context is very difficult to achieve by human and more difficult by machine (Schramm 1971, Bédard 1986). Also, it may be extremely difficult or even impossible to describe the meaning of data in a way that can be appropriately used in the context of interoperability (i.e., interpreted as it was intended and appropriately adapted to the context of interoperability). In human communication, meaning exists only in the mind of the involved sender and of the involved receiver; it results from the matching between the produced/received data and the sender's/receiver's own referents (e.g., education, knowledge, experience, beliefs and context). In fact, a destination agent may wrongly interpret data sent by another agent (i.e., there is a probability to give different meaning from the one intended by the sender). Moreover, the receiver may be uncertain about the appropriate interpretation, and hence, he/she may make wrong assumptions about the intended meaning of data. This may undermine the reuse of data (the main aim of semantic interoperability). In fact, data misinterpretation (i.e., faulty interpretation of data, or uncertainty about its intended meaning) may lead to a poor or inappropriate understanding of real-world phenomena which may cause faulty analysis and bad decisions. The 1) probability of data misinterpretation and 2) the consequences of data misinterpretation constitute **risks to semantic interoperability**.

Such risk of data misinterpretation is more a concern when dealing with geospatial data due to the fact that:

1. Typically, geospatial data are used for real-world phenomena having existence of their own and which are observed, interpreted and represented using models and

³ Notions of *summarizability*, *hyper-cell* and *integrity constraint* will be detailed in chapter 2.

physical descriptions (i.e., data). There is often a gap between the geographic reality and its description. This gap occurs, for example, when vague spatial objects such as air pollution zones, forest stands, soil types and water bodies to name a few (i.e., regions with broad boundaries in the reality) are presented using crisp⁴ geometric representation (e.g., polygons); although they include inherent shape vagueness (Kearns 1997, Worboys and Clementini 2001). Furthermore, several geographic features which may appear as being crisp at first sight (e.g., roads, buildings, and individual trees) end up with uncertain definitions (e.g., width, centerline, beginning and end of roads; exact contour of buildings which differ from the base to the roof; existence of trees which may depend upon height or crown diameter).

2. The Earth's true shape is defined as the geoid which is an equipotential surface that corresponds to the mean sea level. Real-world phenomena are measured with regards to the geoid but are represented on maps via the use of transformations from the geoid to a 3D ellipsoid (which is a mathematical model that most nearly fits the geoid) and to 2D maps (via mathematical models called map projections). Several ellipsoids and map projections exist, leading to different map representations and position of the same feature. In addition, the geoid changes over time as the crust of the Earth changes (as witnessed by GPS systems). The difference between the ellipsoid and the geoid can be of many hundreds meters. Moreover, the projection of real-world phenomena on a map or on a computer display always introduces distortions on some or all of these geometric elements: size, position, angles, distances, area, straightness, etc. (Snyder 1997). Thus, for the same geographic phenomenon on the Earth, we obtain different shapes and positions on different maps. Consequently, geospatial objects represented in a given position on a given datum and according to a given map projection can be wrongly interpreted in another position and according to another projection (e.g., when filtering features based on area, position, distance).

⁴ Having well defined boundaries

3. Geospatial data are typically collected by observing real-world phenomena by humans using their own senses or with artificial sensors. However, the same observed real-world phenomena can be defined differently. For example, a boulevard in one region may be classified as a highway or as a boulevard because the classification criteria are different (this is part of the conceptual uncertainty as defined by Bédard (1988)). Moreover, geospatial environment is continuously changing but not every geographic feature is measured simultaneously. For example, the boundaries of lakes and rivers change over time but lakes and rivers from different regions are not necessarily mapped using aerial photographs or satellite images taken at the same period.

Figure 1.1 shows an Agent 1 communicating the geospatial relation “enclose” to Agent 2. The two agents have their own understanding of what the relation “enclose” means. For the Agent 1, the relation “object A encloses object B” means that the interior of B are completely within the interior of A, the limits of both objects may touch each other. For Agent 2, “object A encloses object B” means that both interior and limit of B are completely within the interior of A. Receiving data from Agent 1, Agent 2 may erroneously think that the intended meaning of the relation “enclose” is identical to his/her/its perception, while it may not be.

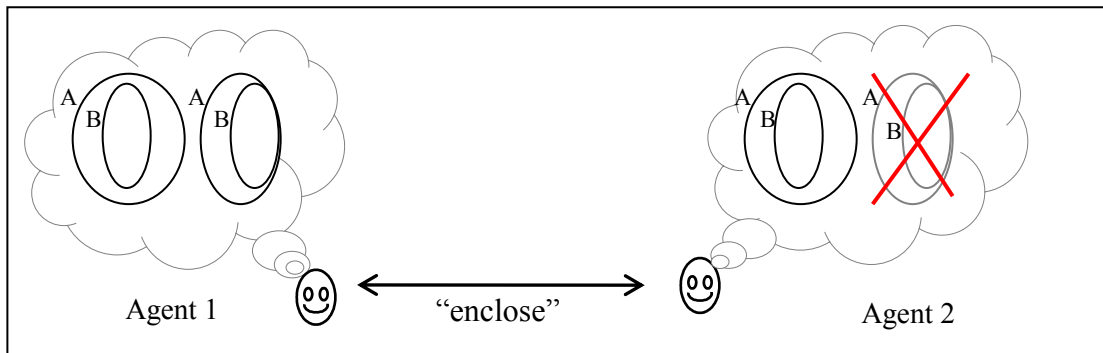


Figure 1.1: Same symbol, different perceptions.

The risk of data misinterpretation is even worse when dealing with geospatial datacubes due to the fact that:

1. in geospatial datacubes, data have undergone complex spatial ETL (Extract, Transform and Load) procedures. This adds to the fact that data are generally collected from other heterogeneous sources having themselves undergone complex procedures (Bédard 1986). During the transformation phase, some interpretations may be formed and several rules, functions and decisions may be applied to the collected data in order to fit business needs (e.g., modifying the terms used to be understood by users in business analysis, extending the boundaries of UTM zones to avoid dividing a large area into two different map coordinate systems but creating important geometric distortions at the same time). All this could make the interpretation of datacubes very difficult.
2. in geospatial datacubes, data are also aggregated or summarized using different methods. This aggregation adds another level of complexity of interpretation since we may need to understand first the method or the pattern used for spatially aggregating data. For example, the buildings surrounded by four streets can be aggregated to form what we call a building block if the density is higher than a given threshold. In order to appropriately interpret the meaning of the resulting polygon, users need to know first the criteria used for the aggregation.
3. geospatial datacubes call for frequent human interventions especially when extracting, cleansing, aggregating, generalizing and integrating geospatial data (Bédard and Han 2008). Such frequency may lead to the modification of data meaning which may cause confusion for end-users and then complicate data interpretation.

Existing approaches of interoperability have tried to deal with the risks related to transactional databases interoperability in a non-systematic way. That is, they are mainly based on agent's experience and not on predefined and ordered criteria. As a result, stakeholders (human users or software agents) have to put extensive time and effort into identifying the risks related to data misinterpretation. Yet, after such intensive practices, they may fail to recognize or even forget to identify the severity of the risks of data misinterpretation and, hence, may inappropriately respond to such risks. Consequently,

existing approaches of interoperability still remain vulnerable to data misinterpretation (faulty interpretation of data or uncertainty about its intended meaning).

The above discussions demonstrate that, while approaches of interoperability between transactional databases can be useful for geospatial datacubes, they present some limits. In fact:

- Existing transactional approaches do not take decisional aspects into account.
- Existing approaches still remain vulnerable to the risk of data misinterpretation and uncertainty. No work has been found that deals with such risks in a systematic way (i.e., its identification, classification, and its treatment).

Based on the limits of existing approaches in dealing with the interoperability between geospatial datacubes, a new approach of interoperability is required. Such approach is needed to categorize the problems related to the heterogeneity and to deal with specific problems related to the interoperability between geospatial datacubes.

Consequently, the general problem addressed by this research is the **absence of a specialized approach that supports the interoperability between geospatial datacubes**. This general problem embraces four specific problems:

1- Absence of concepts and definitions regarding specifically the need for interoperating geospatial datacubes

Many situations justify the need to interoperate geospatial datacubes. However, no study has investigated such need. The questions that may be asked with regards to this point are:

- What is the motivation for the interoperability between geospatial datacubes? In other words, in which situations is the interoperability of heterogeneous geospatial datacubes required? This question will be treated in chapter 3.

- Since data are typically extracted from different transactional sources before being transformed and loaded into a datacube, why shouldn't we interoperate transactional sources and do ETL procedures on the fly (i.e., interpretability between transactional sources, or between transactional sources and datacubes)? This question also will be treated in chapter 3.

2- Absence of a categorization of semantic problems that may occur during the semantic interoperability between geospatial datacubes

Geospatial datacubes may have different data models, different concept definitions, and defined in different contexts. Furthermore, the context in a geospatial datacube which is appropriate and complete for the intended use (i.e., decisional purpose of the datacube), may be less appropriate or incomplete for the application for which the interoperability is carried out. Such semantic differences and context inappropriateness represent major problems to the semantic interoperability between geospatial datacubes. Understanding and categorizing these problems is a crucial step to resolve them. However, we have not found any categorization of semantic heterogeneity in geospatial datacubes. While such categorization would have some similarities with the one of transactional databases, it takes into account the difference in the semantics of aggregation and generalization relationships, the difference in the semantics of summarizing methods and the difference of the semantics of cross-tabulations for every dimension's level and member.

3- Absence of systematic approach that identifies and assesses the risks of data misinterpretation related to the semantic interoperability

The risks of data misinterpretation (i.e., faulty interpretation of data or uncertainty about the appropriate interpretation) can lead to bad outcomes, as they hinder the reuse of data when interoperating geospatial datacubes. Even with such risks, we may still need to interoperate and reuse geospatial datacubes because data definition and collection from scratch can be expensive, or because we need to compare our results to those done by others. Consequently, the risks of data misinterpretation should be identified and assessed when interoperating geospatial datacubes. However, in existing approaches, these risks are dealt

within a non-systematic way. That is, none of the existing solutions of interoperability identified or evaluated explicitly the risk of misinterpretation.

4- Limited support in responding to the risk of data misinterpretation related to semantic interoperability

In order to deal with the risks of data misinterpretation, we need to combine automation and human intervention. This combination is motivated by the fact that it is very difficult or even impossible for the machine to capture and appropriately interpret all possible concepts and relationships that may exist in a particular application (e.g., suggestive concepts and transitive dependencies). For instance, the suggestive concept “Arm of the sea” is used to indicate a sea loch (Wikipedia 2010). Interpreting such concepts requires a common sense and an extensive background knowledge (e.g., knowledge about the organization’s strategy, user’s needs, and legal constraints) which are specific to human (Swanson and Smalheiser 1997). Moreover, machines cannot capture all elements of the context that should be considered in order to determine the “right” interpretation. On the other hand, humans have the capability to achieve a more complete understanding of the context in which geospatial datacube contents have been defined and used. Consequently, the role of humans in the interoperability process is essential and there is almost no indication that they can be completely substituted by technology, at least in the near future. Many approaches have pointed out the importance of human role in interoperating information systems (Visser and Stuckenschmidt 2002, Eklöf et al. 2006, Nagarajan et al. 2006).

However, especially in large scale systems, humans need to be supported by machines as well as machines need to be supported by humans. For example, proposing a set of possible solutions from which they can make a selection, or predefining parts of data matching and propose them to human agents who can make the final decision about whether to accept or reject them (Visser and Stuckenschmidt 2002). In existing approaches of interoperability, the support for human intervention is limited to a set of text input interfaces and requirements in the form of a set of predefined questions (Gruninger and Kopena 2003, Nagarajan et al. 2006). Understanding such texts and responding to such questions may require high skills and considerable efforts (e.g., understand texts, respond to the predefined questions, define and assess the risks, determine a set of solutions, and make a

selection by themselves). However, not all human agents are willing or able to make such effort because either they don't have the knowledge or they don't have time to deal with the risks of data misinterpretation. The situation is even more apparent when dealing with geospatial datacubes due to the fact that:

- Geospatial datacubes generally contain a huge amount of data that cannot be handled by a single person or team.
- Data in geospatial datacubes have undergone complex ETL (Extract, Transform and Load) procedures that lead to additional complexity of data interpretation.
- In geospatial datacubes, data are also aggregated or summarized using different methods. This aggregation adds another level of complexity of data which makes human intervention more difficult.

Consequently, there is a need to provide an approach to facilitate human intervention in the interoperability between geospatial datacubes, and that can be used by both high-skilled and low-skilled human agents to deal with the problems related to the semantic interoperability between geospatial datacubes.

1.3 Hypotheses of the research and objectives

1.3.1 Hypotheses

This thesis is founded on two main hypotheses. First, it is possible to provide an approach that supports the interoperability between geospatial datacubes. Second, it is possible to reduce the risks of geospatial semantic errors during the interoperability of heterogeneous geospatial datacubes. It appeared to us that these two conditions are key for achieving useful or successful interoperability between geospatial datacubes.

1.3.2 Objectives

In order to validate the mentioned hypotheses, the principal objective of this research is to define and develop an approach that supports interoperability between geospatial datacubes while reducing the related risks of semantic incoherence. Although one could do it using

traditional transactional solutions, efficiency can be improved with an enriched approach that specifically supports geospatial datacubes. In chapter 6, we show how the proposed approach, specific to geospatial datacubes, is more effective than existing approaches for transactional databases.

The above discussion (c.f. section 1.2) shows that the definition and the development of an approach to address semantic problems related to geospatial datacubes seems to be beneficial for strategic decision-making. Such an approach requires the definition of semantic conflicts that may arise at various levels of aggregation (cubes, dimensions, hierarchy, levels and measures) when interoperating geospatial datacubes, and the development of a method to support possible human intervention when needed.

This principal objective includes four more specific objectives:

1. To define the semantic interoperability between geospatial datacubes.

We first discuss the need for interoperating geospatial datacubes, and we identify the specific principles of such interoperability. Then, and inspired from existing research works on interoperability, we propose a definition of the semantic interoperability between geospatial datacubes. This objective will be met in chapter 3.

2. To propose a categorization of semantic problems in geospatial datacube interoperability.

The interoperability deals with the problems related to the semantic heterogeneity. In order to develop an approach that supports the interoperability between geospatial datacubes, we first categorize the types of semantic heterogeneity that may occur when interoperating geospatial datacubes. This objective will be met in chapter 3.

3. To propose a systematic approach that identifies and assesses the risk of data misinterpretation related to the semantic interoperability between geospatial datacubes.

We intend to manage the risks of data misinterpretation related to the semantic interoperability between geospatial datacubes (i.e., risks of faulty interpretation of data or

uncertainty about its intended meaning). For that, we need first to identify and evaluate such risks. This objective will be realized in chapters 4 and 5.

4. To propose a method to respond to the risks of data misinterpretation related to semantic interoperability between geospatial datacubes.

We intend to facilitate responding to the risks of data misinterpretation related to the semantic interoperability between geospatial datacubes by proposing a method to assist agents in making appropriate decisions about such risks. This objective will be realized in chapter 5.

1.4 Methodology

This Ph.D. thesis is related to two research axes proposed in the NSERC Industrial Research Chair in Geospatial Databases for Decision Support: the first project aims at developing methods and tools to update on-the-fly geospatial datacubes. The second project aims at creating methods and tools for decisional quality assurance.

From the beginning of this thesis, we noticed that the problem of interoperability of information systems is very broad. Moreover, we did not find any specialized approach that deals with the interoperability between geospatial datacubes. This led us to realize that no single thesis could propose a solution that could achieve perfect interoperability. Consequently, we used a “hypothetico-deductive” approach which consists of proposing a hypothesis to explain some phenomenon (e.g., the interoperability of geospatial datacubes), and developing a prototype to experiment and test this hypothesis. Based on the experimentation result, we can conclude that the hypothesis is either rejected or confirmed. The methodology was based on an iterative process of investigation to clarify the problems and objectives, then to define the theoretical concepts related to the subject of our study. Finally, we develop a prototype to test these concepts.

The process of our methodology includes four concrete phases that will lead to the realization of the above mentioned objectives and the validation of the hypothesis.

Phase 1: Literature review and experimentations to formulate the research problem and objective

In order to better understand the research context and clearly identify the objectives, we carried out an extensive literature review at the beginning of this research. This literature review included definitions of the notions on which we will base our search to support the interoperability between geospatial datacubes such as geospatial databases, geospatial datacubes, interoperability between databases, semantic aspect of interoperability, human communication, standards, ontology and context. Moreover, we reviewed existing approaches for the semantic interoperability of transactional databases such as Bishr (1998), Brodeur (2004), Rodriguez (2000), Kuhn (2005), Hess (2006) and Janowicz et al. (2008). We concluded that, while these approaches can be used to a certain extent to support the interoperability between geospatial datacubes, the efficiency of such interoperability can be improved by developing an approach specific to the interoperability between geospatial datacubes.

Based on this literature review and our initial broader objective, we defined the context of the Ph.D. research and we determined the problems and objectives. Then, we wrote the thesis proposal describing the research context, literature review, problems, objectives, and research methodology. This proposal was defended during the oral and written Ph.D. exams in front of the members of the Ph.D. committee.

Phase 2: Defining the theoretical framework for the research project

We first define a theoretical framework for the semantic interoperability between geospatial datacubes which would contribute to understanding such interoperability and provide a theoretical foundation to better deal with it. In order to define such framework, we first discuss the need for the interoperability between geospatial datacubes. Then, we define the characteristics of semantic aspects of such interoperability. In fact, this interoperability includes geometric and graphic aspects which indicate the semantics of the datacube content. For example, according to its definition within an ontology, a cartographic element may refer to houses according to their roofing (with or without a balcony) in some applications while it may refer to houses according to their foundation in other applications

(see Figure 1.2). Moreover, the interoperability between geospatial datacubes includes semantic information that is defined for strategic decision making (e.g., the semantics of hierarchies, the semantics of aggregation relationships, and the semantics of aggregation functions).

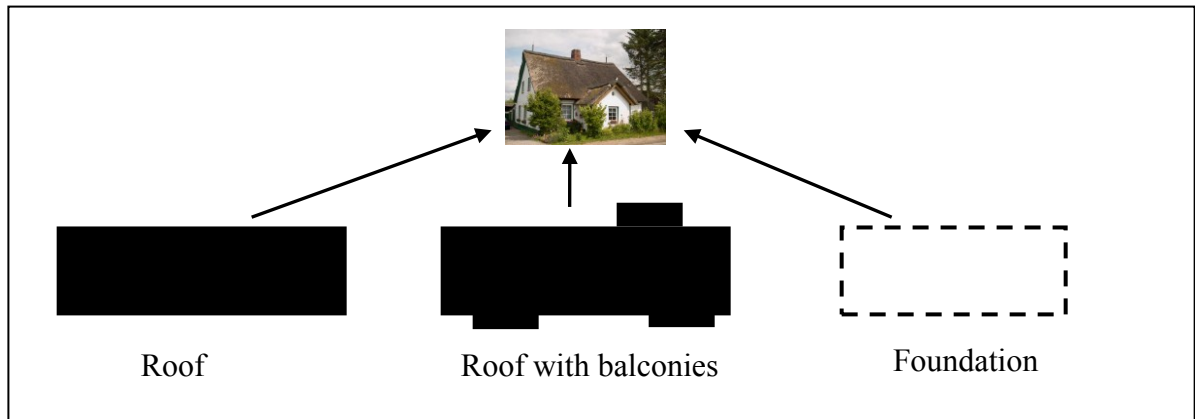


Figure 1.2: Semantics of a polygon in different application.

Afterwards, we define the elements (i.e., data elements and context elements) that should be considered in semantic interoperability. Then, and based on these elements, we categorize the types of the semantic problems that may arise when interoperating geospatial datacubes. Of particular interest were the problems related to the risks of semantic confusion. Finally, we propose a conceptual framework to specifically support the interoperability between geospatial datacubes.

Phase 3: Defining a systematic approach for managing the risks of data misinterpretation related to the semantic interoperability between geospatial datacubes

We first define a systematic approach to define and assess the risks of data misinterpretation related to the semantic interoperability between geospatial datacubes. Our methodology for defining such an approach consists of 1) reviewing the basic notions of risk management in project development, 2) studying existing approaches for managing the risks that affect the use of geospatial data, and 3) determining the causes that may lead to such risks.

Then we propose a framework and an algorithm to support potential stakeholders in dealing with the risks of data misinterpretation. For that, we first review existing methods for

supporting human intervention in the process of interoperability between information systems. Then, we determine a set of elements on which we can base our method (e.g., the external quality of context, the external quality of datacubes schemas and semantic vagueness). Due to the limited time set for realizing our goals in the Ph.D., we focus on two of these elements, i.e., the external quality of context and the external quality of datacubes schemas. This choice is based on four aspects:

- the role the external quality of both schema and context can play in defining the risks of data misinterpretation and determining the severity of such risks,
- the role the external quality of both schema and context can play in supporting a high-level decision-making processes,
- the role the external quality of both schema and context can play in supporting the semantic interoperability process,
- the degree of influence of context quality (e.g., appropriateness and completeness) on the semantics of datacubes.

The proposed method uses different elements of geospatial datacubes as an input, and, as output, it provides 1) a set of indicators to make users aware of the risk of data misinterpretation, and 2) an advice to help users make appropriate decisions to deal with the risks related to the semantic interoperability between geospatial datacubes.

Phase 4: Experimenting with the proposed approach

The proposed approach can then be experimented. Experimentation will show whether our approach is efficient and what other elements we should consider for improving the current research or planning future works. We develop our application in Java environment in order to benefit from its compatibility aspect. Moreover, we can use a number of open source APIs to enhance our application which can be exploited in different environments. The experimentation consists of two main steps:

1. *Implementing and executing the application of the semantic interoperability between geospatial datacubes.* The application consists of three types of agents

communicating together. Two agents represent two geospatial datacubes willing to communicate for a given purpose (e.g., data insertion from one datacube into another). The third agent is responsible for context analysis, and helps the other agents to deal with the risks of data misinterpretation. In order to detect semantically similar concepts of datacubes, we base our implementation on the *GsP* tool developed by Brodeur (2004). The choice of this approach is explained by the fact that 1) the *GsP* tool was successfully tested for supporting the communication process between software agents in geospatial context, 2) the tool was developed with respect to the orientation of the NSERC Industrial Research Chair, and 3) obtaining and understanding the source code was possible thanks to the help of the tool's developer Jean Brodeur; co-director of the current research.

2. *Testing the results.* In this step, we test our approach using an extraction from the content of two different geospatial datacubes developed for different analysis purposes. The first datacube is used to determine the distribution of the population in specific areas and periods (Bernier et al. 2009). The second geospatial datacube aims to analyze and control the forest fire extent. This datacube example is inspired from real statistics that show the risk of fire in Canadian forests according to a set of criteria (e.g., time and regions), and have been published in the SOPFEU's report (SOPFEU 2008). The example is inspired from real cases which consist in determining the risk of forest fire on the Canadian population.

Since we deal with the semantic aspect of interoperability, we use the conceptual models (schema and metadata) associated with the geospatial datacubes to be interoperated. This allows us to avoid technical and structural heterogeneity that may occur between the two geospatial datacubes. We remind that resolving technical and structural conflicts are out of the scope of this project research.

Phase 5: Results analysis

The final phase of this research is to analyze the results of our work. This phase is composed by two main steps:

1. *Reviewing and evaluating the contributions of the thesis*: in order to examine the validity of our hypothesis, we review the results obtained in the previous phases with regard to the defined objectives. Moreover, we review the concepts on which we based our approach and we determine other potential elements to consider in order to enhance our approach. Also, in this phase, we iterate the execution of the developed prototype to check its validity (c.f. chapter 6), and we show the advantages of our contributions. In order to evaluate the contribution of our work, we define a grid that compares the results obtained using the developed prototype with those obtained using the GsP tool.
2. *Drawing the possible perspectives of this work*: In this phase, we show the limits of our approach. Then, based on these limits, we propose some future researches. The future researches aim at defining general instructions to achieve the objectives that cannot be reached in this thesis.

The next UML activity diagram (see Figure 1.3) describes the methodology followed in this thesis:

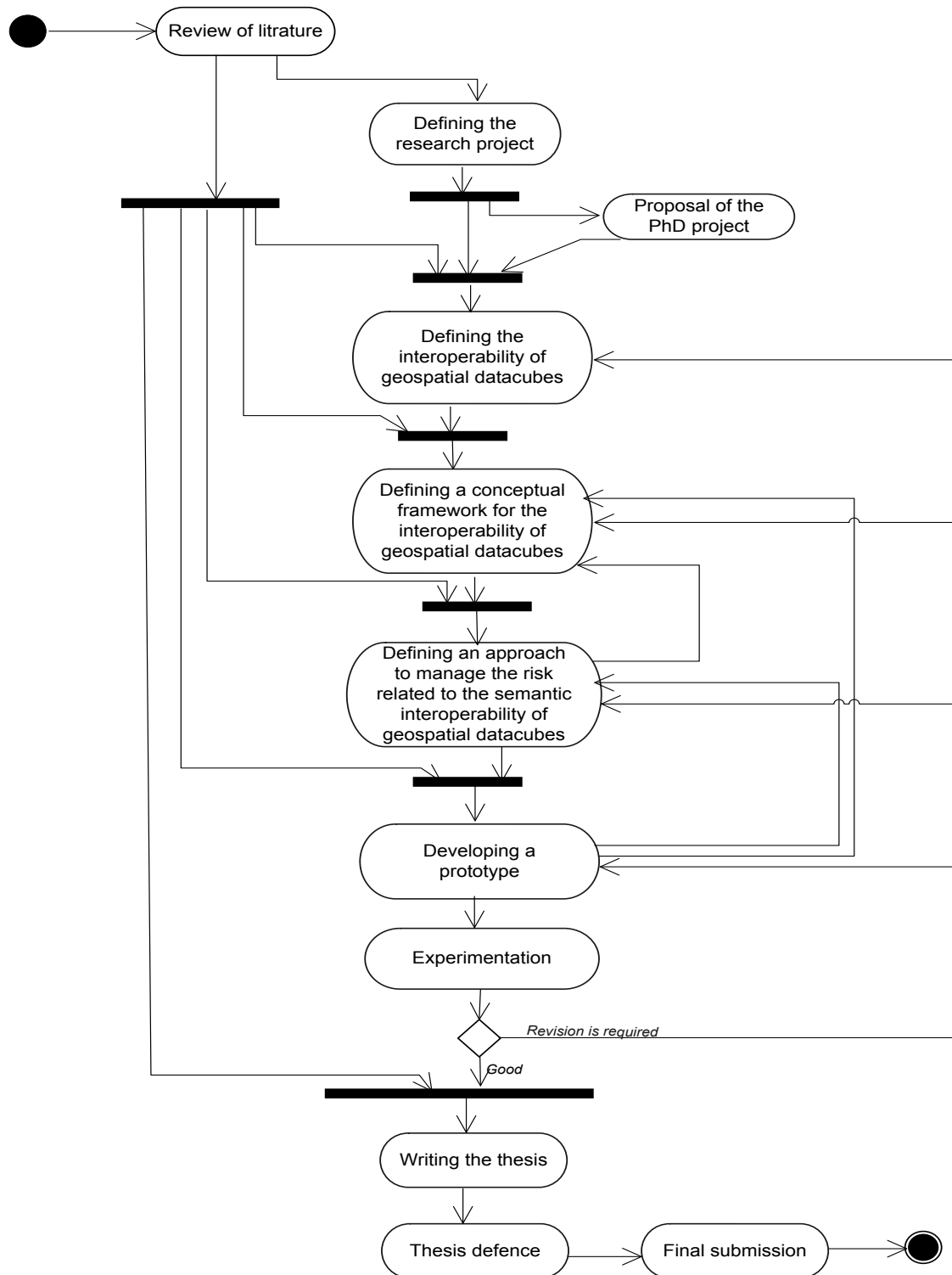


Figure 1.3: UML activity diagram of the research methodology.

We should notice that, some of the results of this thesis (chapters 2 to 5) have been published in the form of nine papers in refereed international conferences and journals. The

results of the chapter 6 are also being adapted for submission in scientific journals. For every paper, I've been the primary author in charge of both the research and the writing.

The published papers in this thesis have been written in collaboration with other co-authors: Professor Yvan Bédard, Associated Professor Jean Brodeur, Professor Thierry Badard, Ph.D. candidate Mehrdad Salehi and professional researcher Sonia Rivest. I am particularly grateful to Yvan Bédard, Jean Brodeur and Thierry Badard for their contribution to the published papers and to the entire research work. Some of these papers, referenced in the bibliography, are available electronically:

- The paper titled "A Conceptual Framework to Support Semantic Interoperability of Geospatial Datacubes" is available at <http://www.springerlink.com/content/e7260103453071h5/> (Sboui et al. 2007).
- The paper titled "Modeling the External Quality of Context to Fine-tune Context Reasoning in Geospatial Interoperability" is available at <http://www.docstoc.com/docs/10977599/Proceedings-of-ARCOE09> (Sboui et al. 2009b).
- The paper titled "Towards a Quantitative Evaluation of Geospatial Metadata Quality in the Context of Semantic Interoperability" is available at <http://bookshop.blackwell.co.uk/jsp/welcome.jsp?source=rss&isbn=1439810125> (Sboui et al. 2009c).
- The paper titled "Une approche basée sur la qualité pour faciliter l'intégration de modèles de cubes de données spatiales" published in the *Revue des Nouvelles Technologies de l'Information* (Sboui et al. 2009d).
- The paper titled "A systematic approach for managing the risk related to semantic interoperability between geospatial datacubes" will appear soon in the *Int. Journal of Agricultural and Environmental IS* (Sboui et al. 2010b).

- A representation of the paper titled “SemEL: A Semantic Model Informed by Cognitive Principles to Support Reasoning about Spatial Data Semantics” is available at <http://www.geoinfo.tuwien.ac.at/lasnavas2010/> (Sbouli et al. 2010c).

In addition, a paper is undergoing redaction with the collaboration of Jean Brodeur, Thierry Badard and Yvan Bédard will be sent to the Enterprise Information Systems (EIS) journal.

We also should notice that we received good comments from the reviewers of our papers, as well as from the audience of the international conferences where we presented our work (e.g., the International Workshop on Semantics and Conceptual Issues in Geographical Information Systems (SeCoGIS-07), and the International Conference on Artificial intelligence (IJCAI-09)).

1.5 Structure of the thesis

This Ph.D. dissertation consists in seven chapters. Next chapter reviews fundamental concepts relevant to this thesis, including geospatial databases, geospatial datacubes and their structure, interoperability of transactional databases, semantic aspect of interoperability, human communication, standards, ontology and context. Chapters 3, 4, 5 and 6 present the contributions of this research. In chapter 3, we explain the need for interoperating geospatial datacubes, we define the characteristics of such interoperability, and we propose a categorization of the semantic heterogeneity that may occur during the interoperability process. In chapter 4, we propose a conceptual framework to support the semantic interoperability between geospatial datacubes. In chapter 5, we propose an approach that allows to define and assess the risks of data misinterpretation related to the interoperability between geospatial datacubes. Also, we propose an approach to help agents to effectively respond to the risks of data misinterpretation by taking appropriate decisions with regards to the problems that may occurs during the interoperability process. In chapter 6, we experiment the proposed approach to show the contribution of this thesis. Chapter 7 draws the conclusions and perspectives of this research.

Chapter 2: Literature Review

2.1 Introduction

This chapter revisits, reviews, and synthesizes the state-of-the-art of several fundamental concepts related to geospatial datacubes and semantic interoperability. First, we review the fundamental concepts of datacubes and their structures, investigate the combination of geospatial data and datacube structure, and study the difference between transactional databases and decisional databases. Second, we study the semantic interoperability of information systems with focus on geospatial databases. Then, we review the basic notions related to the semantic interoperability. Following this, we review existing approaches of semantic interoperability between transactional databases and interoperability between traditional non-spatial datacubes. Finally, we conclude this chapter.

2.2 Datacubes: Characteristics and Functionalities

In order to carry out their functions, managers are engaged in a complex and continuous process of decision-making. The complexity of such process is due to the multiplicity of the alternatives, the high cost of error, and the fast rate of change in different domains. Decision support systems (DSSs) allow many types of decision makers at different organizational levels to systematically analyze problems before making strategic decisions. They provide techniques, models, and tools to identify and solve problems and improve the quality of their decisions. Data warehouses are being considered as efficient components of decision support systems. They are often structured as datacubes, i.e., according to the multidimensional paradigm defined in the Business Intelligence field (BI)⁵.

2.2.1 From Transactional Databases to Datacubes

Transactional databases are widely used in day-to-day operations. They are intended to maintain detailed up-to-date data, with minimum redundancy and maximum integrity. However, strategic decision makers need to analyze, besides detailed data, summarized and

historical data in a rapid way. Meeting this need requires the use of a dual-database approach that supports both daily transactions and strategic decisions needs. Such an approach makes the typical backbone of data warehouses (Bédard and Han 2008). A *data warehouse* is a subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making (Chaudhuri and Dayal 1997). The characteristics of data warehouses are compared to those of transactional databases in Table 2.1.

Transactional Databases	Data Warehouses
Built for transactions	Built for analysis and decisions
Original sources	Copy or read-only data
Detailed data	Aggregate/summary data
Application-oriented	Enterprise-oriented
Current data typically	Current and historic data
Normalized data structure	De-normalized, redundant data structure
Run on DBMS, GIS, Web server, CAD	Run on super RDBMS, MD-DBMS

Table 2.1: Transactional Databases versus Data Warehouses (Bédard and Han 2008).

Generally, a data warehouse is populated by taking data from various legacy sources, and its specific purpose is business decision making, not business operations (Corey and Abbey 1996). Building a data warehouse is a long and complex process, requiring business modeling, and may take many years to succeed. Thus, some organizations choose to develop *data marts*, which contain a subset of data and focus on a given subject.

Data warehouses are often structured according the datacube paradigm (or multidimensional paradigm in the sense of business intelligence) (Gray et al. 1997, Abelló et al. 2006). Moreover, they may also use relational, object-oriented or hybrid implementation models for traditional DBMS or a pure multidimensional model in a multidimensional server. The multidimensional paradigm allows strategic decision makers

⁵ Readers familiar with the concept of datacube may skip Section 2.2.1.

to summarize and aggregate data along different factors in order to have a global view, and to go through the details of each factor and view and visualize the results. Thus, using multidimensional paradigm, data warehouse provides the basis for knowledge discovery (Bédard and Han 2008, Salehi 2009).

In a datacube, analysis is performed according a multidimensional structure, i.e., along a combination of axes of analysis called *dimensions* (e.g., administrative regions, types of vehicles, periods) (see Figure 2.1). Each dimension includes one or several *hierarchies*, each made up of a number of analysis *levels* (e.g., a city-province-country hierarchy has three levels: city, province, and country). An instance of a level is a *member*, for example ‘2008’ is a member of the level “year” of time dimension. Measures such as “number of accidents” are the subject of analysis and are determined according to the members of different levels of dimensions (i.e., the measure is the dependant variable while the members of the dimensions are the independent variables) (Rafanelli 2003). For example, the measure “number of accidents” is determined according to members of the different levels of the dimensions *Type of vehicles*, *Administrative region*, and *Period*. Each combination of measure value(s) with their corresponding members of different dimensions is a *fact*. For example, “the number of accidents of cars in Québec in 2005 is 2000” is a fact. Datacube facts are usually pre-computed in order to speed up query answering (Rivest et al. 2001, Salehi 2009). Salehi (2009) defined the *hyper-cell* concept to describe a model for a number of facts. This concept consists of a pair of a set of levels and a set of measures, where the set of levels includes exactly one level from every dimension in a datacube and is used primarily to define integrity constraints. Another key concept of datacubes is *aggregation* which refers to summarizing measure values by applying aggregation functions according certain dimensions. An aggregation function takes a set of values as an argument and produces a single simple value as the result (Klug 1982, Salehi 2009). Examples of aggregation functions are SUM, AVG, COUNT, MAX, and MIN.

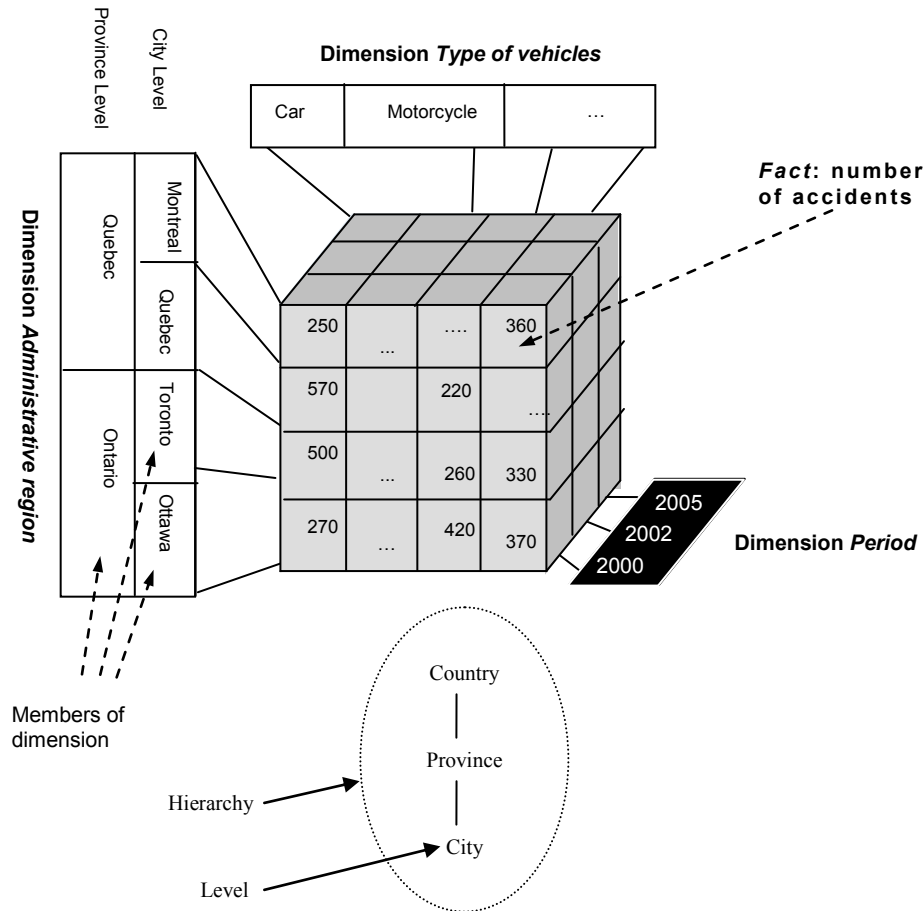


Figure 2.1: An example of a datacube and its elements.

Datacubes can be used by various tools (e.g., as *OLAP* tools, *SOLAP* tools, and *Data mining* tools) for interactive query, knowledge discovery or other purposes. *Data mining* uses different techniques to automatically discover hidden patterns and relationships in large databases and to make predictions (Bédard and Han 2008). *OLAP* software enables decision makers to gain insight into a variety of data views, organized according to a multidimensional structure (Colliat 1995). It provides operators for efficient exploration and analysis of data with a multidimensional user interface. Examples of *OLAP* operators include the following (Pourabbas and Rafanelli 1999, Rafanelli 2003, Malinowski and Zimányi 2008, Salehi 2009):

- *Roll-up*: This operator allows a more general view on data by allowing moving to a higher-level of a member of a given dimension to obtain the corresponding measure. In our example, of the Figure 2.1, by rolling-up from the members Montreal and Quebec of the

level *City* to the level *Province*, we can view the measure “number of accidents” for the province of Quebec. The resulting datacube is represented in Figure 2.2. We should notice that this operator (Roll-up by member) was originally defined in OLAP applications, and that only some of SOLAP tools support this operator. On the other hand, the majority of SOLAP tools support Roll-up by level (i.e., moving to a higher-level of a given dimension).

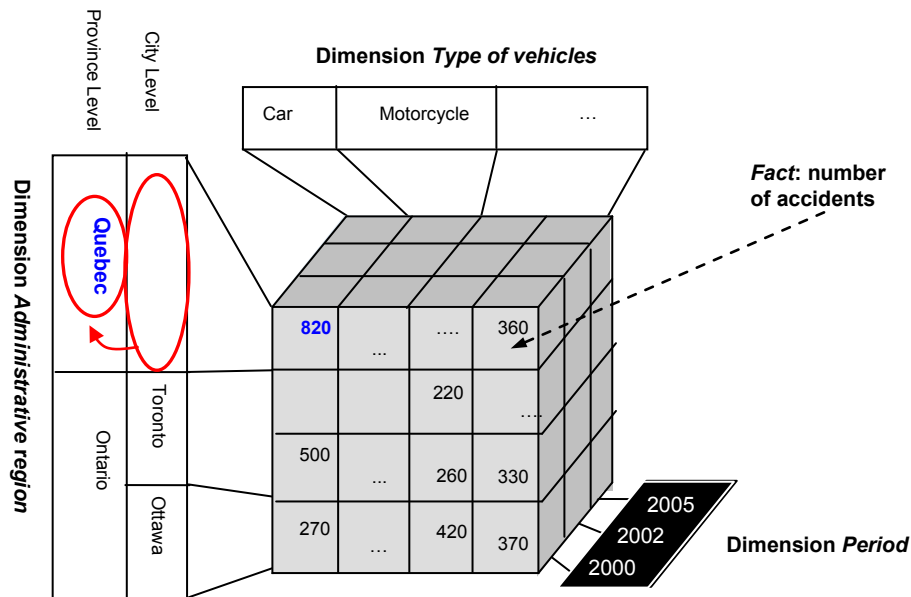


Figure 2.2: The result of applying roll-up on the level *City*.

- *Drill-down*: This operator is the reverse of roll-up. It provides navigation from a higher-level summary to the lower-level detailed data. In our example, of the Figure 2.1, by drilling-down on the member Quebec of the level *Province* of the dimension *Administrative region*, we can view the measure “number of accidents” for all cities of this province instead of viewing it globally for all this province. The resulting datacube is represented in Figure 2.3. We should notice that this operator (Drill-down by member) was originally defined in OLAP applications. The majority of SOLAP tools support both Drill-down by member and Drill-down by level (i.e., moving to a lower-level of a given dimension).

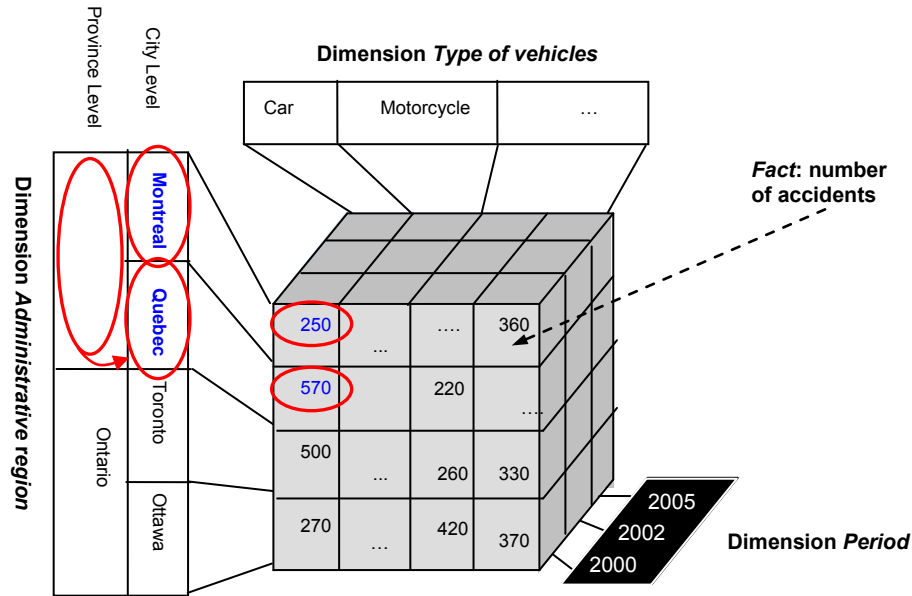


Figure 2.3: The result of applying drill-down on the level Province.

- *Slice* (or Destroy Dimension): This operator returns a sub-cube (or hyper-cell) derived from only a subset of dimensions. For example, slicing the dimension *Type of vehicle* from the datacube results in a hyper-cell with the two dimensions *Administrative region* and *Period* (as shown in Figure 2.4).

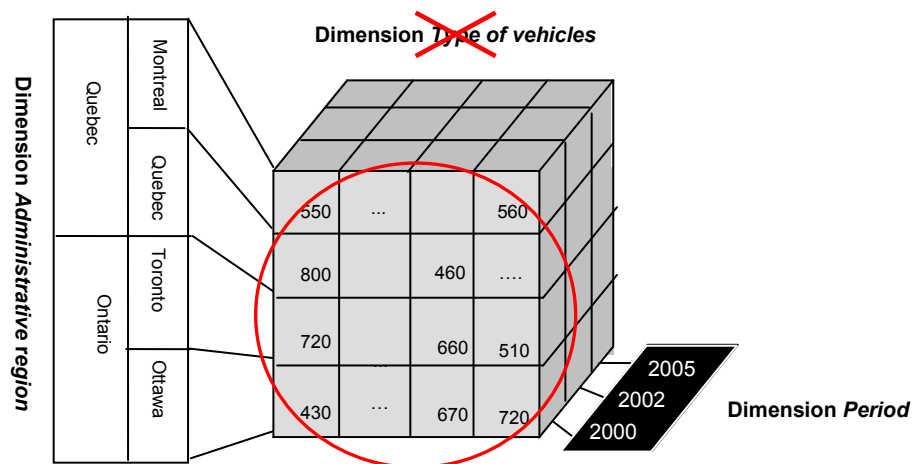


Figure 2.4: The result of slicing the dimension *Type of vehicles*.

- *Dice* (or Restriction): This operator restricts a level's members by removing from it those values that are specified in the operation. For example, dicing Montreal from the level *City* results in the new datacube (see Figure 2.5).

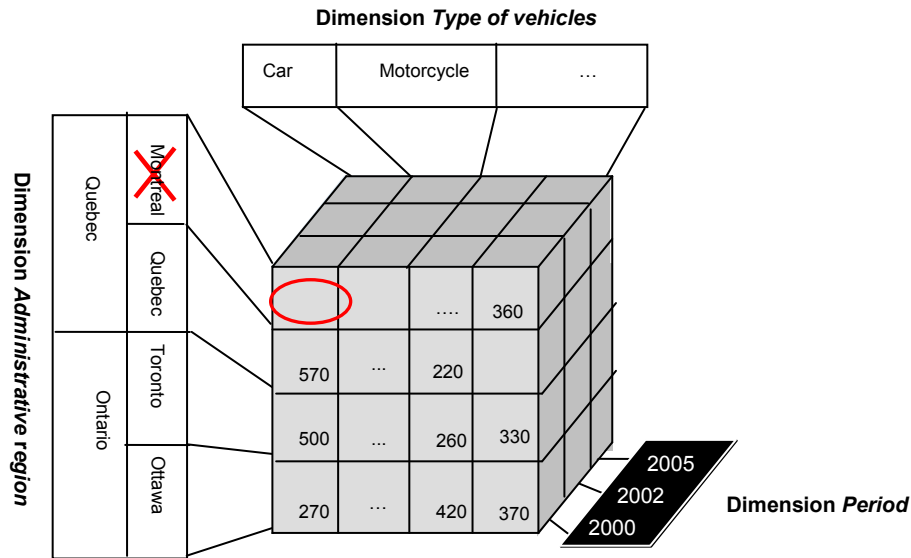


Figure 2.5: The result of applying dice on the member “Montreal”.

- *Pivot* (or Rotate): This operation provides another perspective for a datacube by rotating its axes. For example, applying the pivot operator on the datacube of the Figure 2.1 results in the datacube of the Figure 2.6.

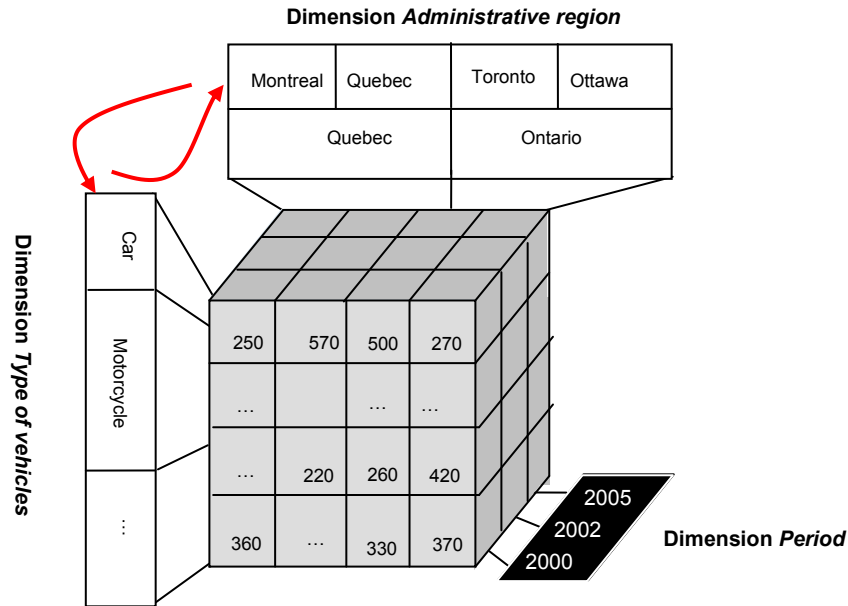


Figure 2.6 : The result of applying pivot on the datacube Figure 2.1.

The implementation of datacubes may take a relational form (Relational On Line Analytical Processing – ROLAP), a multidimensional form (Multidimensional On Line Analytical Processing – MOLAP), or both which called Hybrid OLAP (HOLAP) (Bédard and Han 2008). ROLAP systems use relational database technology and the Structured Query Language (SQL) for storing and querying data. This implementation typically uses three types of models: star schema, snowflake schema, and fact constellation (Pedersen and Jensen 2001). A star schema consists of a single fact table and a table for each dimension (Figure 2.7a). Refining the star schema, by normalizing dimension tables and representing each level of each dimension with one table, leads to the snowflake schema (Figure 2.7b). Finally, the fact constellation is a more complex structure where multiple fact tables share dimensional tables (Figure 2.7c). MOLAP systems store data in a multidimensional array structure rather than a relational database. Finally, HOLAP systems are optimized combinations of the two previous systems (Pendse 2000). They store part of the data in a relational database and the other part in multidimensional arrays.

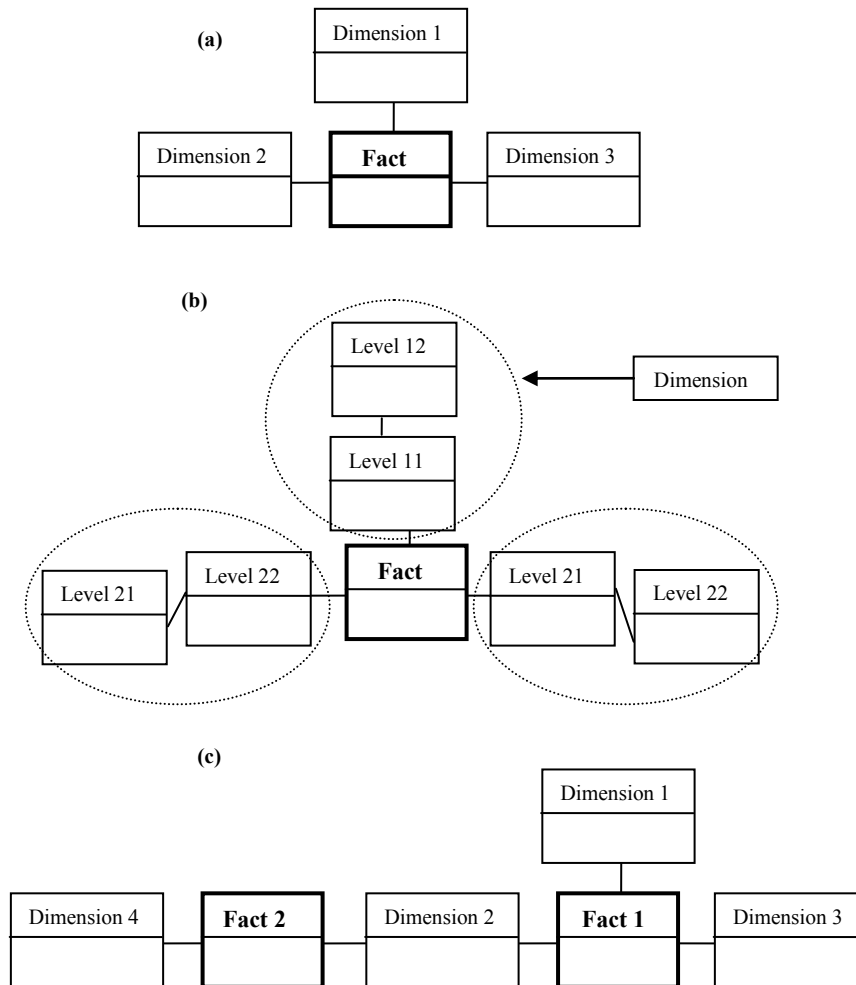


Figure 2.7: A (a) star schema, (b) snowflake schema, and (c) fact constellation schema.

2.2.2 Geospatial datacubes

After revisiting the key concepts of datacube we review the characteristics of geospatial data, and then we discuss the combination of geospatial data with the datacube structure.

Geospatial data, like maps, are very useful for describing and visualizing the phenomena which occupy a location on and beneath the Earth's surface such as buildings, roads, and vehicle accidents. It is widely recognized that geospatial data provide richer information, compared to non-spatial data, about geographic phenomena and facilitate the understanding of these phenomena (Chorley and Haggett 1967, Bilodeau 1991, Chrisman 1997, Goodchild et al. 1999, Bédard et al. 2005, Salehi 2009). The visualisation of geospatial data

allows perceiving relevant information such as geospatial properties of phenomena (e.g., position, shape, direction), their geospatial relationships (e.g., adjacency, distance), their geospatial distribution (e.g., regular, cluster), and their spatiotemporal relationships (e.g., distance during a given period). Perceiving such information helps to draw some conclusions about real-world phenomena which cannot be observed with only nominal information. It has been estimated that up to 80% of all data stored in corporate databases may have a geospatial component (Franklin 1992).

However, geospatial data are more complex in their structures than nominal data. They may be modeled using complex geometry types and complex operations. Geographic Information Systems (GISs) have been recognized as useful tools for dealing with the complex geospatial data in a wide range of disciplines (Goodchild et al. 1999, Dobson 2004). However, GISs have complex interfaces which prevent users from concentrating on data analysis. Instead, they concentrate more on how to get the data. Moreover, GISs are transaction-oriented systems and are not meant to support high-level decision-making processes which need an intuitive overview of geospatial information (Bédard and Han 2008). On the other hand, while non-spatial datacubes are relevant for strategic decision-making, they are not suitable for data visualisation and navigation. They deal with geospatial data like traditional data considering only their nominal aspects.

In order to fully exploit the geospatial component in the context of interactive spatio-temporal exploration and strategic analysis of data, different kinds of tools have been created to integrate geospatial data in a datacube structure. This integration has led to a promising type of decision-support database, known as geospatial datacubes (Rafanelli 2003). Geospatial datacubes extend the datacube concept into the realm of geospatial analysis, geographic knowledge discovery, and geospatial decision-support. They provide capabilities that are not inherent to transaction-oriented systems such as GISs and geospatial database engines. Examples of these capabilities are rapidly summarizing a huge amount of spatiotemporal data and quickly analyzing them at several levels of granularity of space, time, and themes (Bédard and Han 2008, Salehi 2009). Both dimensions and measures of a geospatial datacube may contain geospatial components. There are three types of geospatial dimensions: Non-geometric geospatial dimension, fully geometric

geospatial dimension, and mixed geospatial dimension. We find also two types of measures: numerical measures and geospatial measures (Bédard and Han 2008). Each type may refer to nominal, ordinal, interval, and ratio measures.

Geospatial datacubes have triggered many research topics, such as query processing and spatial OLAP (SOLAP), (Papadias et al. 2002, Rivest et al. 2005, Bédard et al. 2005, Choi and Luk 2008, Malinowski and Zimányi 2008). SOLAP technologies combine GIS and OLAP technologies. They provide users with an efficient geospatial visualisation and navigation and offer several levels of granularity of information allowing them to use their visual capacities to explore data, to get a global picture of geospatial phenomena, and to get more details if needed. This navigation helps enormously in analyzing data and discovering geographic knowledge. SOLAP supports geospatial dimensions (non-geometric, geometric and mixed geospatial dimensions) and geospatial measures (geometric and numeric measures). It allows a rapid navigation through the members of geospatial dimensions on maps using visual variables maintaining the user's flow of thought (Bédard et al. 2005).

2.3 Interoperability: review of key concepts and existing approaches

This section defines the notion of interoperability and presents its different types and levels.

2.3.1 Definition of interoperability

The term interoperability may refer to more than one meaning; it may refer to the openness in the software industry, to the commonality in user interaction, or to the ability to exchange data between systems (Goodchild et al. 1999). The Open Geospatial Consortium (OGC) and ISO/TC 211 defined the interoperability as “the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units.” (OGC 2002).

Geographical interoperability is the ability of information systems to a) freely exchange all kinds of geospatial information about the Earth and about the objects and phenomena on, above, and below its surface, and b) cooperatively run software capable of manipulating

geospatial information over networks (Roehrig 2002, OGC 2002). Bishr (1997) defined the interoperability as the ability of a system to provide information sharing and inter-application co-operative process control. Brodeur and Bédard defined the interoperability as a bi-directional communication between different agents (e.g., providers and end-users) (Brodeur and Bédard 2001). Brodie proposed an elegant definition of interoperability: “Two components (or objects) X and Y can interoperate (are interoperable) if X can send requests for services (or messages) R to Y based on a mutual understanding of R by X and Y, and Y can return responses S to X based on a mutual understanding of S as (respectively) responses to R by X and Y” (Brodie 1992).

The interoperability varies according to the context in which the interoperability occurs as well as the willingness of each entity to achieve interoperability (Goodchild et al. 1999).

2.3.2 Types and levels of interoperability

In the literature there is more than one categorization of the types of interoperability. Goodchild et al. (1999) defined three types of interoperability: technical, semantic, and institutional. The technical interoperability aims to resolve the differences in format, language and user interface. The semantic interoperability focuses on data interpretation (i.e., the ability to interpret data as it was intended, and to adapt data to the context of interoperability). Finally, the institutional interoperability aims at unifying the organisation process and comparing the capacity of institutions participating in the interoperability process. We should notice that technical interoperability may be categorized into structural, syntactic and system interoperability (Bishr 1998, Sheth 1999). In addition, over the last few years, the development of Web services has been remarkable. Web services aim at exchanging the logic of applications over the Web. Hence, we can add to the three types represented above the logical interoperability. As an example of a geospatial Web service, we quote the one developed for the North American Profile of ISO19115:2003 - Metadata (NAP – Metadata) registry in which we participated during the thesis’s process. This Web service allows users to access and explore NAP – Metadata profile over the Web (Brodeur and Danko 2006).

The interoperability may occur at different levels. In table 2.2, eight levels of interoperability were defined (Buehler and McKee 1996, Goodchild et al. 1999, Voisard and Schweppe 1998). The left and right columns of the table show the systems participating in the interoperability. The middle column of the table shows the subject of exchange between the systems (Goodchild et al. 1999).

For the two lowest levels (i.e., network and distributed computing environment), the interoperability was already complete, because these services are interoperable by definition (Goodchild et al. 1999). The remaining six levels still need some work to achieve interoperability. For the tools, middleware, and data some work needs to be done in terms of defining methods and techniques to assist the exchange of services, objects and data. At the application level, we need to define some cooperation and coordination procedures in order to facilitate the comprehension of applications logic and hence assist interoperability. At the two highest levels, we need to solve institutional and social issues related to policy, values, and culture. Interoperability will be more difficult to achieve at the higher levels (Goodchild et al. 1999).

To each of these levels, we have associated the type of interoperability that may be needed. Institutional interoperability is associated with communities and institutions, logical interoperability affects the application level, and technical interoperability is related to the tools, middleware and data. Finally, semantic interoperability affects data stored in different sources. Also, semantic interoperability might affect other levels of interoperability (e.g., logical and institutional interoperability).

A	exchanges	with B	
Information community, institution	policy, values, culture	Information community, institution	Institutional interoperability
Enterprise	agreements, consensus	Enterprise	Logical interoperability
Application	cooperation, coordination	Application	
Tools	services	Tools	Technical interoperability
Middleware	distributed objects	Middleware	
Data store	data	Data store	Semantic interoperability
Distributed computing environment		Distributed computing environment	
Network		Network	

Table 2.2: Levels of interoperability (adapted from (Goodchild et al. 1999)).

In order to support technical, institutional, and logical interoperability, many organisations have defined several standards that aim at unifying the rules and techniques for developing information systems. In geospatial community, Open GIS consortium (OGC), ISO/TC 211, Federal Geographic Data Committee (FGDC)⁶, and other organizations have made important evolution in technical interoperability by defining some standards (e.g., OGC's Geography Markup Language (GML⁷), ISO 19115 Geographic information Metadata (ISO/TC 211 19115, 2003), and FGDC's Standard for Digital Geospatial Metadata⁸). More details about standards will be given in section 2.4.3.5.

Semantic interoperability, which is the main concern of this thesis, goes beyond standardizing techniques. It takes into account the diversity of phenomena representations (i.e., data) and tries to find a way to use data despite their differences in meaning. However, semantic interoperability still faces major problems (Bishr 1998, Harvey et al. 1999, Brodeur 2004, Giunchiglia et al. 2008, Vaccari et al. 2009). These problems are related basically to the semantic heterogeneity. Semantic heterogeneity occurs when there is a difference in the meaning or interpretation of the same or related data. Such difference may be caused basically by 1) the difference in data description, and 2) the lack or the inappropriateness of context information. The difference in data description refers to the fact that data may have different meanings although they are represented similarly (e.g., using the same term to represent two different concepts), or have the same meaning although they are represented differently (e.g., using two different geometric representations to indicate buildings on a map). The lack or the inappropriateness of context information refers to the fact that context information may be missing, or may be inappropriate for the current use (e.g., the lack of information about the precision of data and the geographic scale).

⁶ <http://www.fgdc.gov/>

⁷ <http://www.opengeospatial.org/standards/gml>

⁸ http://www.iso.org/iso/iso_catalogue/

2.4 Semantic interoperability: an analogy with human communication

This section presents the basic notions of semantic interoperability such as semantics, semantic heterogeneity, human communication, ontology and context. Then the section reviews key research on the semantic interoperability of different types of databases and studies their contributions and limitations. The research review is organized according to the types of databases for which the interoperability was defined, i.e., traditional, geospatial databases as well as datacubes.

2.4.1. Semantics of geospatial data

Semantics is the meaning and the interpretation of data. In information systems, semantics is defined as the association between the computer representations of data (e.g., words, geometric primitives) with things in the application domain.

In the context of geospatial data, we believe that geometric and graphic aspects belong to geospatial data semantics. That is, geometric and graphic aspects may convey meanings about geospatial data. For example, a sinusoidal line on a map may indicate a road with several left and right curves. Also, according to its definition within an ontology, a cartographic element may refer to houses according to their roofing in some applications while it may refer to houses according to their foundation in other applications (see Figure 2.8).

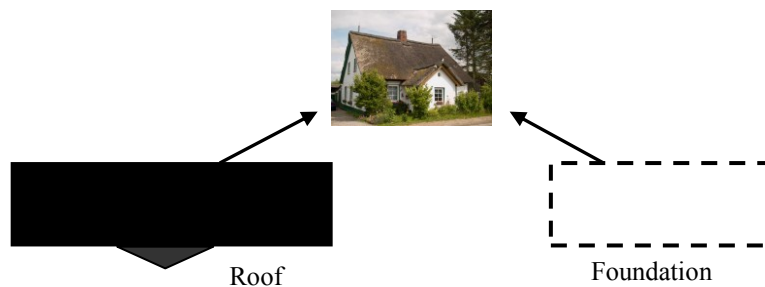


Figure 2.8: Semantics of a polygon in different application.

Also, on a map, polygons may indicate houses, whereas points may indicate light poles (when they are gray) or fire hydrants (when they are black), see Figure 2.9. The fact that

these two types of object were represented with different geometries and graphic visual variables would help us to deduce when visualising the map that an object represented by a polygon is a house, an object represented by a gray point is a pole, and an object represented by a black point is a fire hydrant.

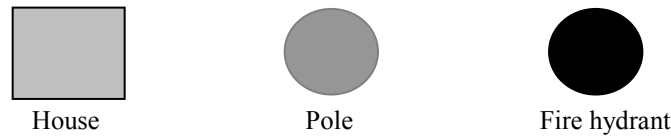


Figure 2.9: Polygon indicating different objects according to their graphic form (shape).

Moreover, geometric and graphic aspects are not inherent to objects but defined according to the needs of a given application. For example, a polygon representing a building may correspond to the roof and may be measured using photogrammetry for a given application, while it may correspond to the foundations and measured using land surveying for another application.

2.4.2. Semantic heterogeneity: a barrier for interoperability

Data heterogeneity represents a major challenge for enabling interoperability between information systems (Bishr 1998, Brodeur 2004). In databases, heterogeneity can be categorized into technical heterogeneity, structural heterogeneity and semantic heterogeneity (Chatterjee and Segev 1992, Denk and Oropallo 2002). Technical heterogeneity involves differences in hardware, operating systems, in database management systems, and in database models (e.g., relational and object-oriented). Structural heterogeneity basically includes differences in data types and data formats (Denk and Oropallo 2002). Semantic heterogeneity, while there is no agreement about its definition (Cercone et al. 1990), generally refers to the difference in meaning, interpretation, or use of the same or related data (Sheth and Larson 1990, Goh 1997, Park and Ram 2004).

2.4.2.1. Semantic heterogeneity in traditional databases

The semantics of databases is usually captured at the time of designing their database model (conceptual schema and metadata) (Hartmann and Link, 2009). Accordingly, in

databases, semantics heterogeneity refers to the difference in describing schema and metadata elements. Examples of such heterogeneity occur when data 1) have different meanings although they are represented similarly (e.g., using the same term to represent two different schema elements), or 2) have the same meaning although they are represented differently (e.g., using two different geometric representations to indicate buildings on a map).

We should notice that semantic heterogeneity of conceptual models does not include differences in modeling techniques used to design or represent databases schemas (e.g., UML, entity/relationship (ER)). This has been the aim of the Object Management Group (OMG) in order to reduce complexity of database designs, and promote their interoperability. The OMG is accomplishing this aim through the introduction of the Model Driven Architecture (MDA). The MDA is a framework for software development that consists of separating the specification of system functionalities from the specification of the implementation of those functionalities on a specific technology platform (Miller and Mukerji 2003). The MDA development process consists of four steps: (1) creating a computation independent model (CIM), (2) creating a platform independent model (PIM), (3) creating a platform specific model (PSM), and (4) generating the code (Blane 2005).

2.4.2.2. Semantic heterogeneity in geospatial databases

The issue of semantic heterogeneity appears more complex when dealing with geospatial data (Gaumond 1998). In fact, in addition to the semantic heterogeneity problems presented above (i.e., semantic heterogeneity of traditional databases), semantic heterogeneity of geospatial data may arise when there is a difference in geospatial characteristics (e.g., geometry, graphics). That is, since geometric and graphic aspects may convey meanings about geospatial data, any difference in the geometry or the graphic aspect (e.g., size, shape, value) is indeed considered as a semantic heterogeneity of geospatial data.

The problem of semantic heterogeneity comes from “the semantics of geometry” and differences in context related to the production of geospatial data (e.g., reference system, projection, and precision). Differences in geometric semantics include several aspects such as variations in the meaning of position, of shape, of orientation, of minimal size and of the

choice and shape of the geometric primitives (point vs. polygon, gravitational centroid vs. geometric centroid, single vs. polyline, detailed line vs. smooth line, quasi-symbolic polygon vs. detailed exact polygon, etc.). Differences in context include dissimilarities in referencing systems (e.g., North American Datum of 1927 (NAD 27) and North American Datum of 1983 (NAD 83)), differences in topological relations (e.g., the related relations “cross” and “intersect” are described by different words, and inversely different GIS systems use the same name for different relationships), different map scales or measurement resolutions, vague and precise spatial concepts (e.g., “close” vs. “next to”), etc. Such differences in the meaning of geometric primitives, spatial relationships and geometric properties of geographic features do indeed have an impact on the meaning of the feature (e.g., trees are defined using their diameter or their height in the topographic database of different provinces in Canada and the minimal dimensions used to define “if a tree is a tree” may also vary).

The causes of heterogeneity in geospatial may also be due to the fact that:

- Geospatial data are often heterogeneous by nature because 1) they are created to represent real world objects which can be observed differently. For example, a boulevard in one administrative region may be considered as a highway in another. Moreover, 2) geospatial environment is continuously changing. For example, the boundaries of a lake may change over time and the update frequency of different sources may differ, meaning that the boundary of Lake A is a 2008 polygon while the boundary of Lake B is a 2010 polygon, or similarly that Lake A boundary is a high-level water Spring polygon while Lake B boundary is a low-level water Summer polygon (because the polygons were not measured simultaneously). It appears that the context of geometry measurement also has impacts on the meaning of geometric primitives.
- Geospatial data never fit together or with the reality because they are goal-oriented models of this reality. Consequently, models eliminate details by simplifying the representation result or process (e.g., representing 3D phenomena in 2D maps) and by focusing on the intended information. Thus, a model is born from a semantic exercise right from the start: what to include, to

which degree of detail, with which level of quality, etc. For example, the Earth's true shape is first defined as the geoid which is an equipotential surface that corresponds to the mean sea level. Then, for simplicity, we use ellipsoids which are mathematical models that approximate sufficiently the geoid in selected areas of the world, or of the Earth as a whole. This also allows more stable measurements over time since the geoid varies over time while the selected ellipsoid doesn't. The difference between the ellipsoid and the geoid can be many meters in altitude and a few milli-degrees in the direction of the local vertical axes. Moreover, the projection of real-world phenomena on a 2D map or on a 2D computer display cannot be done without distortion, either of angles, areas or of both at the same time (Snyder 1997). Thus, for the same phenomenon measured on the Earth with regards to a geoid and then related to the same ellipsoid, we obtain different shapes and positions when we represent it on different maps made with different projections. These differences may be up to hundreds of meters (Bédard et al. 2005). Figure 2.10 illustrates the geoid-ellipsoid-map transfer process

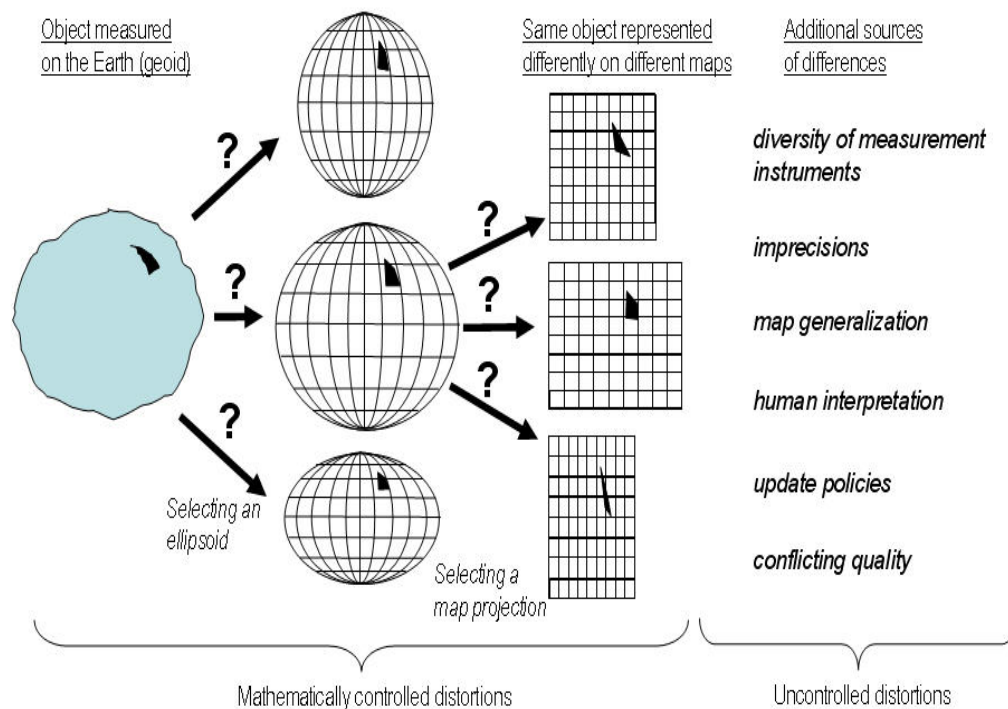


Figure 2.10: The geoid-ellipsoid-map transfer process (Bédard 2005).

- Geospatial representation always results from generalization. In the generalization process, different types of modification can be carried out (e.g., some objects and categories of objects can be eliminated or replaced by a new symbol, some objects can be displaced, some shapes can be simplified, some topological relationships can be changed, and some groups such as “building blocks” can replace individual buildings where the density is too high (Bédard et al. 2005)). Consequently, the same real-world phenomena may be represented differently depending on what type of generalization is carried out. Figure 2.11 shows an example of the geospatial detailed-generalization mismatch where detailed data represent more faithfully real-world phenomena but make the map unreadable while generalized data produce a readable map but unfaithful representation (Bédard et al. 2005). Thus, data may be interpreted differently. For example, Figure 2.11 (a2) is interpreted as a block with 11 buildings, while Figure 2.11 (b2) is interpreted as a block with 8 buildings.

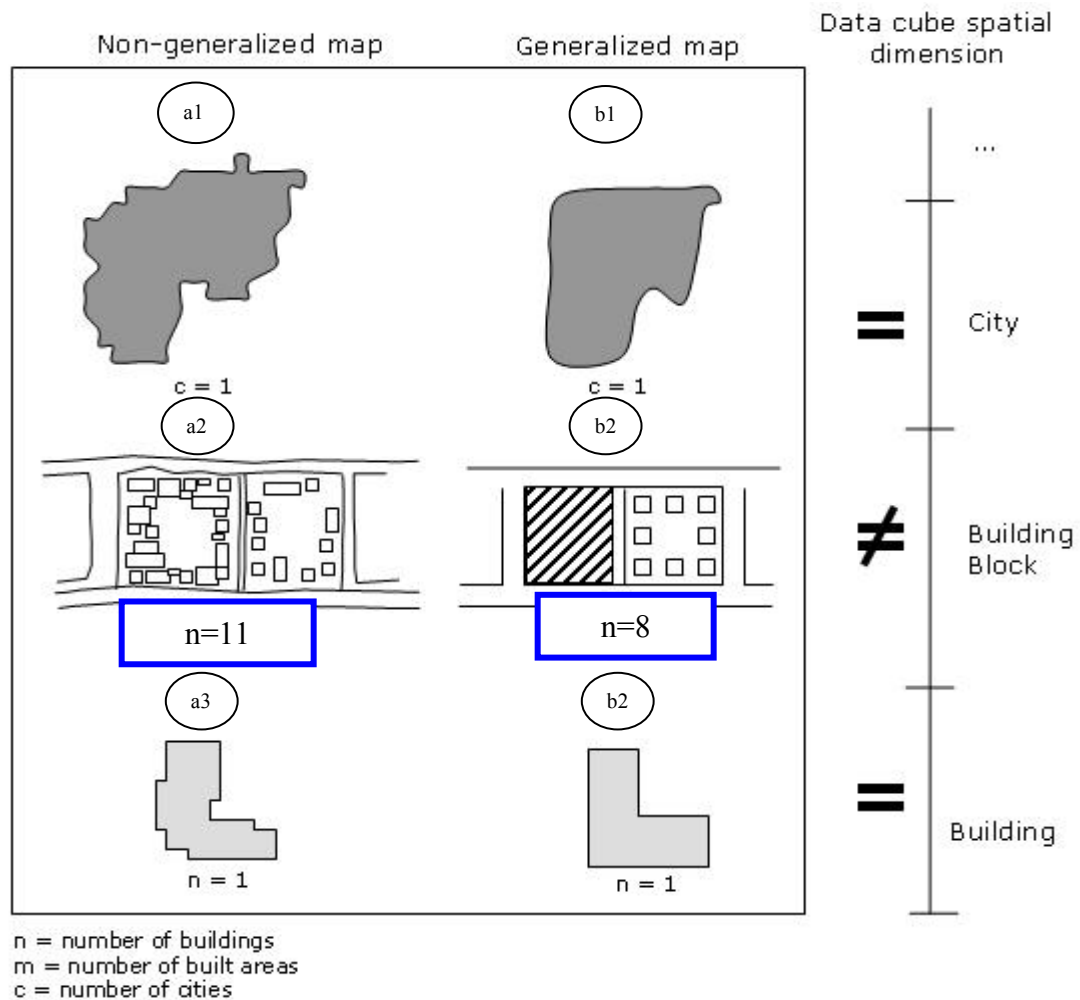


Figure 2.11: Example of the spatial aggregation-generalization mismatch (Bédard et al. 2005).

- Geospatial data specifications as well as their acquisition techniques evolve over time (e.g., use of new instruments, methods and referencing systems).
- Different disciplines conceptualize geographic space in different ways (Dumedah 2005). For example, a health organization may conceptualize a forest as a space that sustains the life of insects spreading the *West Nile* virus. Whereas, the municipality may conceptualize the forest as a space for recreational activities. Such differences will have impacts on the boundaries as well as the existence of forests in the two datasets.

- Geospatial data may be represented according to different referencing systems. For example, the same real-world phenomenon may be represented according to an x-y-z coordinate system from a 3D digital terrain model, or to an x-y coordinate system based on a map projection. The “meaning” of the geometry representing objects differs between datasets (i.e., the semantics of their geometry varies).
- Geospatial concepts may be described using different models. For example, a bridge may be represented as 1 or 2 dimensional object depending on geometric or semantic criteria (e.g., polygonal if larger than 2 lanes, linear otherwise; polygonal if public, linear if private).

2.4.2.3. Semantic Heterogeneity in traditional (non-spatial) datacubes

We have found very few works that discussed the semantic heterogeneity in traditional datacubes (Bruckner et al. 2001, Frank and Chen 2005). Bruckner et al. (2001) identified four types of semantic heterogeneity that may arise when integrating distributed data warehouse systems:

1. Dimensions having similar schemas, but belonging to different datacubes.
2. Different levels of detail for equivalent dimensions (e.g., time dimension has two different sets of levels in two different datacubes: month, year in datacube 1, and day, month, year in datacube 2).
3. Levels with the same name but different meanings (e.g., two levels with the same name “Province”; one developed in France and another developed in Canada).
4. Different level names but the same meaning (e.g., dimension level *Week* in English, *Semaine* in French).

While this categorization considered some semantic conflicts in the dimensions belonging to different datacubes (e.g., different levels of details), it does not pay attention to the difference of measures nor members. Moreover, this categorization does not take into account the difference between metadata associated to the elements of datacubes.

Frank and Chen (2005) defined three categories of semantic heterogeneities:

1. Cube-to-cube heterogeneity: this kind of heterogeneity occurs when we use different dimensional models to create semantically related multidimensional databases (e.g., star vs. snowflake schema).
2. Dimension-to-dimension heterogeneity: this kind of conflicts arises when the semantically related dimensions have some differences in:
 - Dimension schema: these differences occur when two data cubes have different dimension hierarchies, with possibly different dimension levels.
 - Dimension members: these differences arise when two cubes have different members corresponding to the same level in their semantically related dimensions.
 - Naming differences: these differences occur when mismatched names were used for semantically related dimensions or semantically related levels of different datacubes.
3. Measure-to-measure heterogeneity: it occurs when measures in different datacubes have different names, different values, different formats, or different units. Consequently, this heterogeneity may be further categorized into:
 - Measure naming difference which occurs when different names were used for semantically related measures of different datacubes.
 - Measure value difference which arises when different datacubes have semantically related measures with different values.
 - Measure scaling difference which occurs when different semantically related measures have different units.

This categorization investigated some semantic conflicts in heterogeneous elements of datacubes (i.e., cube, dimension, level, and measure). However, the investigation was

limited to common conflicts caused by schema difference (e.g., naming mismatch), and by some few examples of metadata elements (e.g., unit of measure). Also, we noticed that the cube-to-cube semantic heterogeneity is defined at the implementation level (e.g., star vs. snowflake schema). We believe that implementation differences should not be considered at the semantic level since the semantics of a datacube is independent from the implementation details. Moreover, the categorization did not take into account the difference of hierarchies belonging to different of datacubes.

We should notice that we have not found any categorization of semantic heterogeneity of geospatial datacubes. While the previous categorizations of semantic heterogeneity provide relevant notions to be used to explore the semantic heterogeneity in geospatial datacubes, they did not pay attention to all possible elements of datacubes (e.g., hierarchy). Moreover, although metadata is an important element to consider in any semantic heterogeneity investigation, the previous categorizations merely took into account the difference between metadata associated with datacubes content. Metadata are especially important for geospatial datacubes since nobody can estimate the real meaning of a geospatial measure, facts' geometry and members' geometry without knowing the spatial reference system used for measures, facts' and members' geometries and the whole datacube. Consequently, we still need a more complete categorization that investigates semantic conflicts that may occur in geospatial datacubes.

2.4.3. Key notions of semantic interoperability

Several researchers have been working to define approaches to deal with the issue of data heterogeneity. They have based their work on some key notions such as human communication, ontology, context, metadata and standards. In this section we review these key notions.

2.4.3.1 Human communication and interoperability

We have witnessed the development of the multi-agent systems paradigm that intends to imitate human behaviour and intelligence. Since people communicate easily to exchange data and share services, multi-agent systems can be used to simulate the communication

process between humans. Some researchers have based their work on the communication process between humans in order to support interoperability between different information systems (Uitermark et al. 1999, Brodeur and Bédard 2001, Xhu and Lee 2002).

The communication process is based on the representation of a real-world model evolving in human's mind (i.e., mental model) using signs (i.e., terms) and on assigning meanings to those signs (semiotics). In the remainder of this section, we present the human communication process and the notions of mental model and semiotics. We should note, however, that discussing these notions in detail is not the focus of this thesis.

Human communication process:

Human communication consists of a source, a sign, a communication channel, a destination, a possible source of noise, and feedback. These concepts have been accepted and used for the last 50 years (i.e., following the publication of Shannon's book about the theory of information.) (Shannon 1948). Shannon defined the communication process as the act of reproducing at one point either exactly or approximately a message selected at another point. Many researchers evaluated the strength and weakness of Shannon's model (Schramm 1971, Denes and Pinson 1993, Bédard 1986, Poore and Chrisman 2006). Some of them have enriched Shannon's model with cognitive elements that are absent from Shannon's machine-to-machine communication (Schramm 1971, Denes and Pinson 1993, Bédard 1986, Brodeur 2004).

Schramm (1971) defined a model to depict the communication process (see Figure 2.12). In the first stage of this process, the source has a model of real-world that evolves in his/her mind (i.e., a mental model). This model is encoded to a set of signs which do not have any meaning. The signs are transmitted to the receiver who decodes and interpret them (i.e., gives a meaning to the signs). The receivers base their interpretation of messages on their knowledge and experience. However, people have different knowledge and experience. That's why, during a communication process, data (i.e., signs) may be interpreted differently by different receivers.

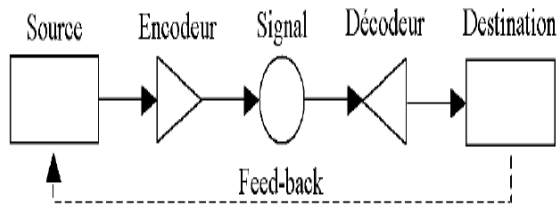


Figure 2.12: Communication process (Schramm 1971).

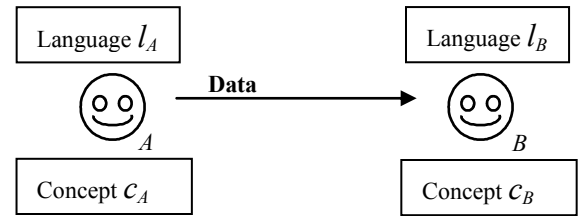


Figure 2.13: communication between two agents A et B.

Figure 2.13 shows an agent A who wants to communicate a concept c_A described with a language l_A to create data (i.e., signs). Agent B receives data and decodes them according to his/her personal vocabulary (i.e., l_B) to create the information corresponding to his/her interpretation to the data (i.e., c_B). Since l_A is more or less different from l_B , c_B will be more or less different from c_A (Denes and Pinson 1993, Reenskaug et al. 1995).

Moreover, in order to communicate, agents need to have a common background (commonness), a concept introduced by Schramm (1971). Without a common background, an agent would have a hard time to correctly interpret a message. Schramm used the concept “field of experience”, to determine if a received message would be interpreted as it was intended by the source (see Figure 2.14). For example, data to be sent refer to a set of coordinates according to the North American Datum of 1927 (NAD 27). In order to correctly interpret these coordinates, receiver needs to be familiar with this datum.

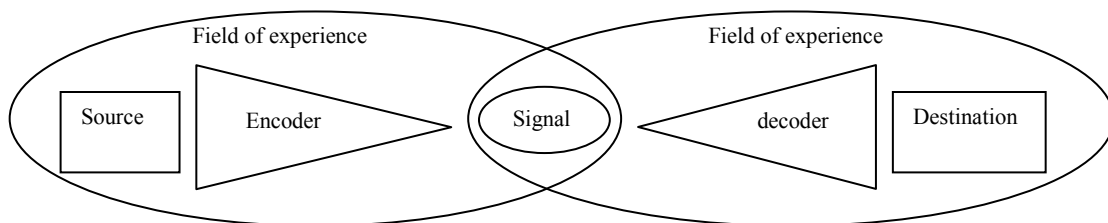


Figure 2.14: Schramm's concepts.

Mental model

Humans continuously try to organize and explain some impressions of the world. Gazzaniga, a brain scientist, thinks that there is an interpreter in the brain which tries to understand and reason about these impressions (Gazzaniga 1988). In order to understand

and reason, human brain creates an idea of the world around us (i.e., mental model⁹). Humans rely on this model to act and react in the world (Norman 1983, Reenskaug et al. 1995). The mental model differs from one person to another. For example, forest fire may be modeled in different ways: a fireman may model it in terms of fire severity, forest stand and forest surface. Whereas, a responsible for the population safety may model it in terms of insulation, wind orientation, and position relative to neighbouring homes.

Mental model and geovisualization

The geovisualization (or cartographic visualization) is a discipline emerging from the fields of cartography, GIS, and visualization (MacEachren 1995, MacEachren and Kraak 1997, Chrisman 1997, Rhyne et al. 2006). In fact, with the emergence of interactive technologies, graphical interfaces and decision support systems, the role of cartography was extended to include techniques for exploration and reasoning (Beaulieu 2009). That is, instead of data production (the aim of traditional cartography), geovisualization stresses the construction of a mental model through interaction and exploration of data.

In order to explicitly represent mental models using a set of signs, one needs to assign meanings to those signs (i.e., semiotics).

Semiotics

The science of the relation between signs and their meaning is called semiotics. From a philosophical point of view, one could claim that the meaning is intrinsic to the object (i.e., a Kantian approach) while someone else could claim that such meaning exists only in the observer's mind and isn't inherent to the object. Although it is beyond the scope of this thesis to debate over such philosophical issues of semantics, we admit that our approach relies on a non-Kantian approach where meaning is the result of an interpretation (and consequently isn't intrinsic to objects). In this sense, Ogden and Richards (1923) depicted the relation between signs and meaning in what is called the meaning triangle which represents the relationship between the sign and the object being signed (see Figure 2.15).

⁹ The concept of mental model was introduced by Kenneth Craik in his book titled "The nature of explanation" (Craik 1943).

The sign references a concept which points to the object. This concept has to be specified explicitly and unambiguously to enable an assignment between the sign and the object (Janowicz 2003). In the context of the interoperability between systems, this specification has to be machine-readable (ontology aim – cf. section 2.4.3.2).

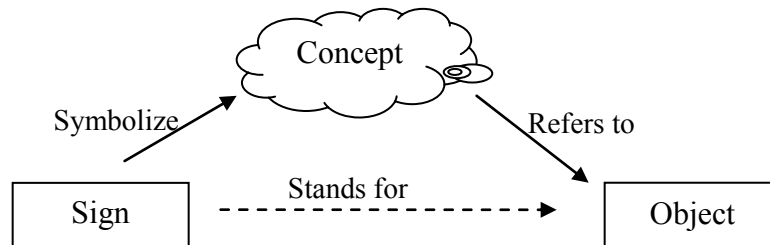


Figure 2.15: Meaning triangle based on Ogden & Richards (Ogden and Richards 1923).

The study of semiotics is closely related to the notion of ontology; since semiotics is the study of signs and ontologies attempt to do exactly that, represent meaning.

2.4.3.2 Ontology

The term Ontology was borrowed from philosophy, it refers to the science of describing the kinds of entities in the world and how they are related (Smith et al. 2004). In computer-processing, many definitions were proposed. These definitions range from simple specifications of a concept set, to the logical theories providing the intended meaning of a given vocabulary.

Gruber (1993) defined an ontology as an explicit specification of a conceptualization. That is, an ontology is a description of a set of concepts and the relationships that can exist between them. In the context of communication, Gruber (1993) used the term ontological commitments which are agreements between a set of agents to use a shared vocabulary in order to communicate about a domain of discourse (e.g., classes, relations, functions, or other objects). He considers conceptualization as a set of extensional elements that exist in some area of interest (i.e., the set of objects, concepts, and other entities and the relationships between them) (Genesereth and Nilsson 1987). Guarino (1998), while agreeing with Gruber that the ontology is an explicit specification of a conceptualization, considers the conceptualization as a set of explicit assumptions about the intended meaning

of a set of terms (i.e., vocabulary). These assumptions have usually the form of first-order logical theory which has the vocabulary as predicate names.

Maedche and Staab (2001) consider ontologies as (meta) data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine-processable semantics. Sowa (2000) considers the ontology as the categories of things that exist or may exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D.

An ontology represents the concepts according to various levels of abstraction. It can be classified into three types (Guarino 1998, Gomez-Perez 1999, Rifaieh 2004, Semy et al. 2004):

- *Top-level ontology*: Top-level or Upper-level ontology describes general concepts. These concepts are independent of a specific domain. An upper-level ontology is often characterized as representing common sense concepts, i.e., those that are basic for human understanding of the world (e.g., space, time, event, action, issue, etc).
- *Domain ontology*: it provides vocabulary about the concepts and the relationships that depict the theories of a given domain. (e.g., a medical ontology representing the concepts related to the medical field).
- *Application ontology*: describes the concepts specific to a particular application (e.g., an ontology of the management of the urban transport of a city). For geospatial databases, an ontology of application can be a conceptual model, a data dictionary or a specification of a geographical product. For example, the specification of the National Topographic Data Base of Canada (NTDB) contains an ontology which describes the concepts which represent a set of topographic phenomena (Brodeur 2004).

Ontologies provide a common understanding of a domain among humans, among systems and between humans and systems, and then facilitating the sharing of knowledge (Sugumaran and Storey 2002, Castano et al. 2005). Many researchers have based their works of interoperability on the notion of ontology (Bishr 1997, Hakimpour 2003, Daconta

et al. 2003, Obrst 2003, Brodeur 2004, Castano et al. 2005, and others). Bishr (2007) used ontology to create a federated model that would assist the integration of different databases. Brodeur et al. (2003) proposed five ontological facets of geospatial data interoperability which consists of the different facets of reality that may occur in the context of geospatial data interoperability. The facets are: (1) the reality itself, (2) the cognitive model of the reality, (3) the set of conceptual representations, (4) the databases' internal representation of reality, and (5) the conceptual representations.

Ontologies may have different forms (e.g., taxonomy, conceptual model and logical theory). The efficiency of any interoperability process depends on the ability of each form to process semantics. Structured languages (e.g., logic-based language) enable to represent rich semantics (Daconta et al. 2003, Obrst 2003).

We should keep in mind the difference between logical theories and conceptual models. In fact, conceptual models aim at modeling data that will be saved in a database for a specific reason. Generally, conceptual model language (e.g., UML (2010)) does not support automatic interpretation. Whereas, logic based languages (e.g., OWL) have a high expressivity and contain logic expressions that support automatic reasoning and interpretation by providing various logical statements (e.g., equivalent, inverse, and transitive relations) (OMG-IBM 2003). For example, Ontology Web Language (OWL) is a language for defining and instantiating Web ontologies in a way that can be understood and handled by machine. It supports the representation and the use of knowledge in a distributed environment such as the Web (Bechhofer et al. 2004). For instance, the following logical statement uses the expressions `equivalentClass` and `subClassOf` to define the concept "Province":

```
<owl:Class rdf:ID="Province">
  <rdfs:subClassOf rdf:resource="#AdministrativeRegion" />
  <owl:equivalentClass rdf:resource="#Territory"/>
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="#Québec" />
    <owl:Thing rdf:about="#Ontario" />
  </owl:oneOf>
</owl:Class>
```

Ontologies play an important role in enabling semantic interoperability between information systems by providing a common understanding of a set of concepts among information systems enabling them to share data. However, ontologies do not usually include all necessary elements of context.

2.4.3.3 Context

People communicate by exchanging messages between them. These messages implicitly or explicitly include “circumstances” that are related to the situation of the sender, or to the message to be sent (i.e., the context). In order to interpret the sender’s message, the receiver should be aware of the context related to the message.

The notion of context has more than one definition. Brézillon and Pomerol (2001) defined the context as a “shared knowledge space that is explored and exploited by participants during the interaction”. Sciore et al. (1994a) defined the context as a piece of data to be the metadata relating to its meaning, properties (e.g., its source, quality, and precision), and organization.

Context has been granted special attention by different researchers in the realm of databases interoperability:

- For Sciore et al. (1994b), context can be used to define the meaning, content, organization, and properties of data. They proposed an extension to SQL, which they called Context-SQL. The extension can be used by users who are interested in modifying their contexts dynamically during the query process.
- Yu et al. (1991) used context to determine common concepts characterizing the similarities between attributes in multiple databases.
- Sheth and Kashyap (1992) considered that each object has its own context which refers to the context in which a particular semantic similarity is defined. This context may be similar or different from the context in which the object was defined. As a consequence, two objects which are semantically closer in one context may be semantically different in another one. For example, a “province” and

“territory” are semantically closer in the Canadian context (i.e., “province” and “territory” belong to the same administration level); however, they are semantically different in the French context (i.e., “province” and “territory” are hierarchically in different administration levels).

- Ouksel and Sheth (1999) considered the context as the primary vehicle to capture the semantic of objects during the interoperability process. For example, the context can help to distinguish between whether the term “cricket” refers to an insect or a sports game.
- Brodeur (2004) considers the context as the union of all the intrinsic and extrinsic properties of geographic phenomena.

Context is used in communication in order to restrict the situation humans are talking about (Vieira et al. 2005). In the computational environment, context includes the set of relevant information that approximates as closely as possible the intended interpretation of data (e.g., techniques used to measure geospatial data, methods used for aggregating geometries, user’s profiles, and psychological and social aspects that surround data collection). Consequently, context plays a crucial role in interpreting data in semantic interoperability. In geospatial domain, context helps users to locate and understand geospatial data, to properly assess the meaning and the fidelity of the presented geometries, and to evaluate their fitness for specific purposes. It provides relevant information related to the circumstances in which data have been defined (e.g., techniques used to define phenomena, accessibility, and methods used for aggregating geospatial data) and can be used. Such information can bear some meaning and, hence, help agents to capture the semantics of data.

However, context still remains a key issue in semantic interoperability. Two major concerns may be distinguished; the first concern is whether any definition of all the “circumstances” is possible. The second concern is the uncertainty about the completeness and appropriateness of context information to the purpose of the interoperability. In fact, context information which is appropriate for a specific application may be less appropriate for another: for example, the dates of certain photographs displayed on Google Earth are

often one year old, which may have no impact on several usages but may mislead others. In fact, the quality of context information may be insufficient for certain data reuse. Poor context information quality may cause a risk of misinterpreting data, and may undermine the reuse of geospatial data (Agumya and Hunter 2002), i.e., the main aim of interoperability. The **appropriateness of context information for a given application can be referred to as the external quality (fitness-for-use)**. Being aware of the fitness-for-use, stakeholders may take precautionary measures when interpreting data. Consequently, the fitness-for-use of context information seems important when interpreting data.

In this regard, we distinguish two kinds of context: *production context* and *use context*. Production context is any information that can be specified by the data producer (e.g., the method of data collection, referencing system used to distribute the data). On the other hand, characteristics that surround user's application (e.g., referencing system and scale) are considered as use context. Contexts can be thematic (e.g., data acquisition method), geospatial (e.g., geospatial referencing system used) or temporal (e.g., the time of data acquisition).

Both production context and use context can be explicit or implicit. Explicit context is directly specified, whereas implicit context is not specified and can be derived from explicit context (Schmidt et al. 1999). Typically, in database realm, the **production context is typically represented explicitly as metadata**.

2.4.3.4 Metadata

Metadata include comprehensive information about data since they may indicate the circumstances in which data was produced and how data must be used (the context of intended use). They may partially describe the data, and may also include content-independent information like location and time of creation. Examples of geospatial metadata include the content standard for digital geospatial metadata (CSDGM) defined by the FGDC, and the ISO/TC 211 Geographic information - Metadata (ISO/TC 211 2003).

Metadata descriptions present three advantages:

1. They enable the representation of details about data such as the format and organization of data. We should notice here that metadata is essential, especially in technical interoperability since it is not feasible to derive information about technical aspects (e.g., a file format).
2. They enable representation of domain knowledge describing the information domain to which the underlying data belongs. This knowledge may then be used to make inferences about the underlying data (Sheth 1999).
3. They help users to locate and understand information and indicate to them how data should be used. Moreover, metadata helps users evaluating its fitness of data for specific purposes.

Metadata play a crucial role in the interoperability since they provide the information related to context of geospatial objects (e.g., techniques used to define geospatial objects, quality, and accessibility). Such information can convey some meaning and hence facilitating the understanding and the exchange of geospatial data during the interoperability process among humans, among systems and between humans and systems.

But, metadata is usually defined in different formats (textual, tabular, chart etc.). Some standards define a common way to identify and represent metadata (e.g., ISO 19115, FGDC's Content Standard for Digital Geospatial Metadata) to facilitate their understanding and their exchange between different information systems.

2.4.3.5 Standards

Standards have been widely used in a variety of domains and applications in order to assist the interoperability between information systems. The main aim of standards is to provide a common description of data. Using a common description, systems will have less difficulty interpreting data and exchanging it between them (Albrecht 1999, Fonseca et al. 2002 a). Examples of geospatial standards include the Geography Markup Language (GML) which is based on the Extensible Markup Language (XML¹⁰), and defined by the Open Geospatial Consortium (OGC). Also, ISO/TC 211, an international body for standardisation, defined

some standards related to geospatial data such as ISO/TC 211 Geographic information - Metadata standard (ISO/TC 211 2003), which defines a common set of metadata terminology, definitions, and extension procedures.

Many researchers used standards to overcome some problems of heterogeneity between information systems. Badard and Braun (2004) have developed an interoperable platform called OXYGENE that embeds SOAP (Simple Object Access Protocol) (W3C, 2007) and WSDL (Web Service Description Language) (W3C 2001) to support the interoperability of geographic applications on the Web (i.e., geographic Web services).

While standards support the interoperability by proposing common descriptions that can help enormously to overcome technical and structural barriers, they gave a little support to overcome the semantic heterogeneity. In fact:

- 1) in some cases, the more you want to get people to agree, the more differences will be found since there is no single geographic reality that overrides all others (Nyerges 1991, Harvey et al. 1999).
- 2) there still is a lack of widely-accepted standards that deals with all data models that are used by people and organizations (Harvey et al. 1999).
- 3) standards restrict the way humans or systems represent the reality (unless they have a built-in extension mechanism such as “stereotyping” in UML).

Semantic interoperability between geospatial datacubes goes beyond standards, and aims at defining ways to deal with the complex problems generated by the differences that lie behind apparently similar descriptions.

2.4.3.6 Geospatial data quality

Data quality in datacubes is an important topic in strategic decisions. In fact, these data are used as a basis for decision making in large organizations dealing with health care, environmental management, transportation, among others. Any failure in providing data

¹⁰ <http://www.w3.org/TR/REC-xml/>

with acceptable quality may lead to incorrect or bad strategic decisions (Ballou and Tayi 1999, Salehi 2009).

Two main groups of definitions of geospatial data quality can be found in the literature. The first group associates the quality of a product or a service with respect of standards and specifications, allowing to reduce the errors in the product. The second group associates quality with the satisfaction of the users' needs, i.e., a product with a good quality level should meet or exceed the users' needs. These two groups of definitions are commonly identified as "internal quality" and "external quality" (known as the "fitness-for-use") (Chrisman 1983, Kahn and Strong 1998, Aalders 2002, Dassonville et al. 2002, Devillers et al. 2002).

- *Internal Quality*: this quality is generally placed from the point of view of data producers. The internal quality of geospatial data is typically defined by its actuality, geometric and semantic accuracy, its genealogy, logical consistency, and completeness which reflect the producer's perception of quality.
- *External Quality (fitness-for-use)*: external quality is placed from the point of view of users. It corresponds to the concept of adequacy to the user's needs (i.e., the "fitness-for-use") (Juran et al. 1979). Bédard and Vallière (1995) define "the external quality as the set of characteristics which make geospatial data ready to meet user's needs in a given application".

The internal quality can be evaluated by making the comparison of the actual data description with requirements defined by the user or by the producer (David and Fasquel 1997). Generally, the internal data quality is evaluated using the following elements (or indicators): *lineage*, *completeness*, *logical consistency*, *accuracy*, *precision*, and *resolution* (Guptill and Morrison 1995, Azouzi 2000, ISO/TC,211 2002, Aalders 2002, Van Oort 2006). Data quality elements are explained below:

- *Lineage*: it refers to the history of a geographic dataset. It describes the source of data as well as the acquisition and derivation methods including all transformations involved in the data production process (Van Oort 2006).

- *Completeness*: it measures the exhaustiveness of the data in terms of geospatial and thematic properties (Brassel et al. 1995).
- *Logical consistency*: it is the degree of adherence to logical rules of data model, attribution, and relationships (Guptill and Morrison 1995, Van Oort 2006).
- *Positional accuracy*: it describes accuracy of the position of geospatial objects (Guptill and Morrison 1995).
- *Temporal accuracy*: it is the accuracy of the temporal attributes and temporal relationships of features (Van Oort 2006).
- *Attribute accuracy*: it measures the accuracy of quantitative and qualitative values assigned to the thematic attributes (the population of an urban area, the city name, etc.) of the spatial objects.
- *Resolution*: it refers to the level of detail that can be represented.
- *Precision*: it refers to the degree of detail that can be recorded.

Furthermore, data quality may also include data accessibility (Bédard and Vallière 1995), appropriate amount of data, data believability, ease of manipulation, data interpretability, data reputation, and data security (Pipino et al. 2002).

On the other hand, the evaluation of the fitness-for-use of data depends on the viewpoint of its suitability for a specific application. Accordingly, the fitness-for-use generally cannot be objectively described and evaluated by data producers because the same data can be intended for different uses (Bédard and Vallière 1995, Agumya and Hunter 1997, De Bruin et al. 2001). Still some authors used the previously defined elements for the internal quality and proposed ways to help users to evaluate the fitness for data to his/her specific use (Goodchild and Gopal 1990, Devillers 2004). For example, Devillers (2004) proposed an intuitive approach to communicate the information about geospatial data quality to users in order to improve the evaluation of its fitness-for-use.

The elements of internal quality can be used to evaluate the geospatial data imperfections in a geospatial datacube. Such evaluation can support the interoperability process. In fact, for example, receiving data from a source agent, the receiver agent can decide whether to consider the received data or not (or even to continue the communication process or not). A receiver (e.g., an end-user) is more interested in the external quality of data (i.e., its fitness of use). That is, the receiver's interest is figuring out whether the received data is complete for his/her specific use or how much this data is appropriate to his/her need. However, generally data and context associated to it cannot be evaluated prior to its use because the same data can be intended for different uses. Consequently, it is important to evaluate the external quality of data and its context during its use (i.e., after knowing its intended use). While many research works have been carried out to evaluate the external quality of geospatial data (Bédard and Vallière 1995, Agumya and Hunter 1997, De Bruin et al. 2001), no work has been found that evaluates the external quality of context information during the interoperability process. In this research project, we propose to evaluate the external quality of context information during the interoperability between geospatial datacubes.

2.4.3.7 Semantic similarity

In order to deal with the problem of semantic heterogeneity between different data sources, we need to compare the semantic similarity between their content. The semantic similarity is the semantic relation which indicates the degree of synonymy between the two concepts. The concepts are generally represented within an ontology and the semantic similarity is typically determined according to the position of each concept within a common ontology.

Many researchers have been interested in measuring the semantic similarity between geospatial concepts to support the interoperability. For example, Brodeur (2004) defined the Geosemantic proximity notion (*GsP*) which allows to determine qualitatively the geosemantic similarity of geospatial concepts. Also, Rodriguez (2000) proposed the Matching Distance model to define the semantic similarity between geospatial object classes. The smaller the semantic distance between concepts, the closer the concepts are in meaning.

Semantic similarity is generally related to the notion of ontology. However, in our project, we argue that the semantic similarity is more than comparing different concepts within an ontology. Indeed, **it is also comparing other semantic aspects of the data to be exchanged, such as the fitness-for-use of data and the fitness-for-use of the context related to such data.** These kinds of information are generally not included in ontologies.

2.4.3.8 Perspective of the basic notions

In the previous sub-sections (2.4.3.1 - 2.4.3.7), we reviewed key concepts related to the interoperability of information systems (human communication, ontologies, context, metadata, standards, data quality, and semantic similarity).

Ontologies provide a common vocabulary in a specific domain or application. This vocabulary can be used by agents participating in a communication process to exchange and interpret messages. Hence, ontologies can effectively support the interoperability process. Nevertheless, ontologies should be attached to the context in which the elements of ontologies (i.e., concepts, their properties and relationships) are defined and used. Defining explicitly the context in which geospatial, temporal and thematic data are stored would support machine understanding and hence guide the communication process between different geospatial information systems. Metadata enriches the context and helps agents to evaluate the relevance of geospatial, temporal and thematic data to be exchanged during the interoperability process. Both data and context/metadata may be incomplete or inappropriate to the use for which the interoperability is carried out. We believe that embedding relevant information about data and context (e.g., their quality) can help enormously to define the semantics of the data exchanged between agents, and hence, to enable better interpretation of data during the interoperability process. Hence, embedding relevant information about data and context would fine-tune the semantic interoperability between databases in general and between geospatial datacubes in particular.

2.5 Semantic interoperability – review of related works

Semantic interoperability has been a major focus of the information systems research community due to the high rate of technological change, the lack of accepted standards, the

autonomy of the information systems, etc. (Fileto 2001, Nienaber 2008). In the last two decades, we have witnessed a growing interest in the semantic interoperability of geospatial databases. In contrast, there have been very few works dealing with datacubes. To date, there has been no work, to our knowledge, on the semantic interoperability between geospatial datacubes.

This section reviews key research on the semantic interoperability between different types of databases and studies their contributions and limitations. The review is organized according to the type of databases they are defined for, i.e., traditional and geospatial databases as well as datacubes. Since we deal with geospatial datacubes, we will give more attention to geospatial databases and datacubes.

We should notice that, in many research works, semantic interoperability overlaps with other notions such as semantic integration and information retrieval. These notions are also concerned with overcoming the semantic conflicts that may occur between different sources. Accordingly, although we focus on key works on semantic interoperability, some works on semantic integration and information retrieval can be found in our review (c.f., sections 2.5.1 – 2.5.3).

2.5.1 Semantic interoperability of traditional databases

The aim of most of the approaches for traditional databases was to define a common way to access and use different databases as if they were only one. The approaches can be classified as follows:

1. *Global schema approach*. This approach consists of creating a global schema for a set of local databases. Each user's application is provided with its own view of the global schema. It is very similar to the design of a conceptual schema for a set of applications in a single DBMS environment, where each application in such environment is provided with its own view of data (Breitbart 1990). For example, Frank and Chen (2005) proposed a global schema approach to integrate heterogeneous databases with different schema structures.

2. *Federated approach.* This approach does not require the creation of a global schema. For each application, the database administrator creates a schema of the data that the database has agreed to share with other local databases. It aims at defining a common access to different databases which have different structures and different DBMSs (Hammer and McLeod 1980, Breitbart 1990, Sheth 1999). For example, Sheth defined a five level architecture (see Figure 2.16): At the lowest level, we find the local schema of each database. This schema may be represented using different formalisms (e.g., E/R, UML). At the second level, each schema is translated to component schema which conforms to the federation schema. At the next level, we find the export schema which defines the available data in the database. This schema enables data filtration in order to manage the transactions in the database. The integration of the different export schemas allows to create the federated schema. At the fifth level, we find the external schemas which consist of the specification of data needed for specific users or applications.

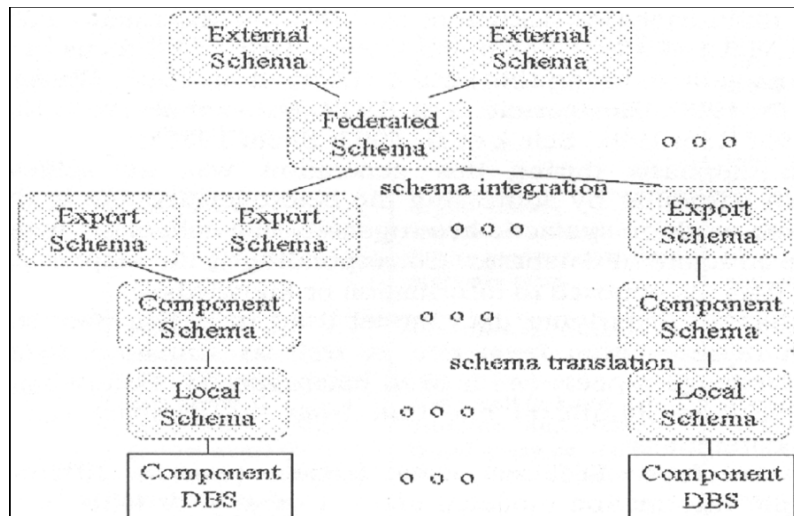


Figure 2.16: Data federation architecture (Sheth 1999).

3. *Multi-database query language approach.* Users may need to interrogate/query multiple databases that have different schemas. This need is apparent when users (Krishnamurthy et al. 1991):
- have the same objective in interrogating different databases,
 - need to formulate queries about the content of databases,

- would like to have a unified view of all the databases, or
- would like to be able to update all the databases through a unique view.

Krishnamurthy et al. (1991) proposed an approach that provides a unified view of different databases and a common language for interrogating this view. They defined a language called Interoperable Database Language (IDL) which can be used to query heterogeneous databases.

In these approaches, ontologies play a key role in enabling such interoperability. Matching ontologies of different databases is a plausible solution to the semantic heterogeneity problem (Vaccari et al. 2009). Ontology matching can be considered as an operation that takes two graph-like structures, such as conceptual model descriptions, and generates a set of semantic links between the nodes of the ontologies (Giunchiglia et al. 2008). These links can be used to translate or mediate schema elements and hence enables different databases to interoperate (Vaccari et al. 2009).

2.5.2 Semantic interoperability of geospatial databases

Enabling semantic interoperability between geospatial databases is more challenging than the one between traditional databases. This is because the issue of semantic heterogeneity is more difficult to deal with when handling geospatial data (cf. subsection 2.4.2.2).

In the last two decades, semantic interoperability of geospatial data has received a growing attention. Several attempts have been made to automatically overcome semantic heterogeneity of geospatial data, notably the Semantic Formal Data Structure model (Bishr 1997), the Matching Distance model (Rodriguez 2000), the Isis approach (Benslimane 2001), geosemantic proximity (Brodeur 2004), the G-Match tool (Hess 2006) and the similarity-based information retrieval approach (Janowicz et al. 2008). In these approaches, ontology played a key role in supporting the interoperability between geospatial databases.

Bishr (1998) extended the Formal Data Structure proposed by Molenaar (1994) to reconcile heterogeneous representations of geospatial data. He suggested the representation of cognitive semantics (i.e., common interest for a group of individuals) as a set of rules and

proposed the formalization of semantically similar classes in independent contexts. The rules aim at associating real world facts to objects in the Discipline Perception World (i.e., the domain of abstraction of the real world). These are heuristic rules that were originated in an expert's past experiences (i.e., human knowledge). That is, geometric and thematic data, and knowledge can be encapsulated into an abstract object type to facilitate the interpretation of real world objects. This is known as data/knowledge packets that aim to represent both knowledge and data in a unified model (Doyle and Kerschberg 1991). These packets use triples of the form (Operation, Object, Meaning) where: operations are functions, procedures, or constraints, objects are database entities and attributes, meaning is the corresponding abstracted real world fact.

Rodriguez (2000) proposed an approach to identify semantically similar object classes belonging to different databases based on their semantic distance. The approach consists of identifying and analyzing the difference and resemblance of class definitions, functions, and attributes of different classes. Class descriptions are specified using an ontology. The evaluation process is based on semantic neighbourhoods where subclass-superclass relations and partOf relations between concepts are represented by directed arcs in an ontology. However, the matching distance model does not take into account geometric properties of the object classes.

Brodeur (2004) considered that the semantic interoperability of geospatial data involves a bi-directional communication process between agents who use their background to interpret data and real world objects. He defined a conceptual framework for geospatial data interoperability based on human communication. In order to measure the semantic similarity between geospatial concepts, he defined the notion of *geosemantic proximity* (*GsP*) which consists of evaluating the semantic similarity between geospatial concepts (i.e., similarities between their intrinsic and extrinsic properties). Intrinsic and extrinsic properties are derived from geospatial database object's properties, operations and associations and are specified within different ontologies.

Fonseca et al. (2002b) proposed an ontology-driven geographic information system (ODGIS) model that supports the interoperability between different geospatial sources. The OGDIGIS model enables multiple data sources with different geospatial and temporal

characteristics to interact. The architecture for the ODGIS includes an ontology editor and an ontology translator as well as a user interface to facilitate the navigation within a given ontology as well as between different ontologies.

Nambiar et al. (2006) developed an interoperability framework within the GEOscience Network (GEON)¹¹ project to facilitate geospatial information retrieval. The framework allows a data provider to register a geographic dataset with one or more mediation ontologies using the GEON registration procedure. This procedure consists of allowing users to manually define a semantic relationship between some concepts in a given ontology (source ontology) to some concepts in another (target ontology). Consequently, users are able to navigate between different ontologies for which a match is defined. Then, different datasets can be queried in a similar way (Nambiar et al. 2006). However, users may have to formulate multiple queries to find relevant data. Moreover, the proposed framework requires that users adhere to a registered community.

Hess et al. (2006) proposed a geo-ontology-based approach to support the interoperability between geospatial data. They developed an algorithm and an implementation of a geographic ontology matcher (*G-Match*) to define a similarity measure between two different geographic ontologies. The algorithm considers the features of different concepts (name, attributes, taxonomy, conventional and topological relationships) and proposes some weights for each geographical feature to compute the similarity between the different concepts. However, there is not a perfect combination of weight factors to each concept features. Such combination requires a specific adaptation of the weight, depending on the input ontology.

Lutz and Klien (2006) presented an ontology based approach to support the retrieval of geographic information. The approach proposed a query language and a graphical user interface to allow a requester to formulate a query from which an ontology is derived. The ontology is then used to search a catalogue for a data source that provides all the information required to answer the requester's query. However, in the proposed approach, it is assumed that a requester searches for only one source at a time. Moreover, as the data

¹¹ <http://www.geongrid.org>

quality (e.g., resolution and precision of data) and the purpose for which data are defined can vary widely, there is no explanation of how the representation of domain and application ontologies has to be adapted for a specific purpose.

Janowicz et al. (2008) proposed a similarity-based information retrieval system to support users and systems to retrieve needed information. They assumed that the needed information is represented as individuals (or features) and concepts (or feature types) which are used as the basis for information retrieval. The approach consists of mapping user's representation to features and feature types provided in an ontology (called the geospatial data infrastructures (SDI)). Also, the authors provided use cases for a human Web interface, as well as for an integration workflow. However, the retrieval system requires the user to define the search and context features manually. Defining such features is not an easy task especially when these features are not available within an ontology.

Recently, Vaccari et al. (2009) proposed an approach that uses domain ontologies to integrate geo-services. The approach used the structure preserving semantic matching (SPSM) ontology matcher defined by Giunchiglia et al. (2008) as a solution to the semantic heterogeneity problem between different implementations of required geo-services in Spatial Data Infrastructure (SDI) domain. Then, they applied the matching algorithm to resolve the semantic heterogeneity problem scenario in peer-to-peer (P2P) infrastructure, i.e., without any central control. In this approaches, peers share explicit knowledge of "interaction models" in which they are engaged. However, to run an interaction model, a peer should know which interaction model it wants to execute and with which peers it will be interacting.

Existing approaches tried to solve the semantic problems that may occur during the interoperability process of different sources (traditional and geospatial databases). They have been based on ontologies to represent concepts, their properties, and their relations with other concepts. However, these approaches do not stress the importance of dealing with the semantics characteristics of datacubes (e.g., semantics of aggregation and generalization and the semantics of summarizing methods and algorithms).

2.5.3 Semantic interoperability of datacubes

Although the semantic interoperability between databases has attracted the attention of many researchers, there have been few works dedicated to datacubes.

Mangisengi et al. (2001) proposed a framework to support the interoperability of heterogeneous datacubes based on the federated approach. Moreover, they used a mediator to manage user's queries. The mediator receives sub-queries from the federated layer and translates them into the query language of the local datacube. The proposed approach uses the benefits of XML as a standardized, universal format for data exchange. However, the approach focused on solving structural heterogeneity and did not propose ways to deal with semantic conflicts of heterogeneous datacubes.

Bruckner et al. (2001) defined a framework for integrating information stored in distributed data warehouses. They used the topic maps paradigm (ISO JTC 1/SC 34, 2008) to represent data stored in data warehouses. The framework uses XTM (XML Topic Maps) (Pepper and Moore 2001) in order to describe schematic mapping between topic maps of local data warehouse resources and integrating them to global topic maps (see Figure 2.17). However, semantic conflicts are left to end-user to deal with. The authors proposed only some outlines about how some specific semantic conflicts can be dealt with (e.g., defining a global-year topic to solve the conflict between Gregorian year and a Chinese year).

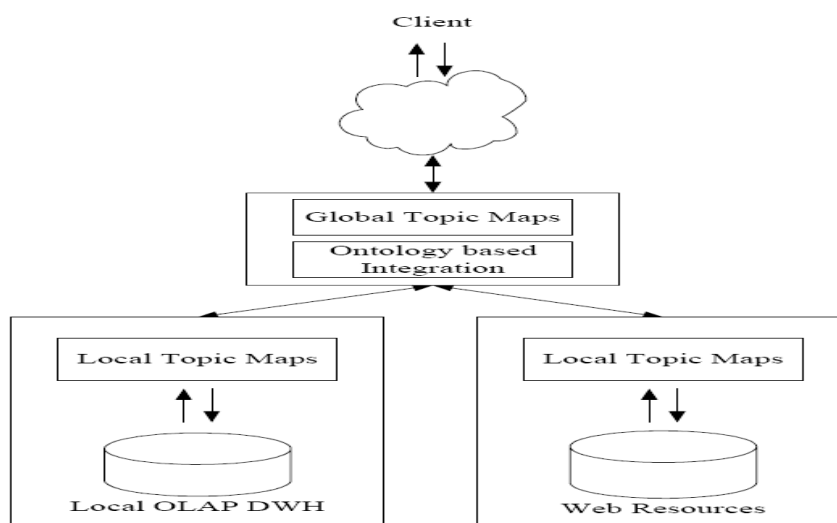


Figure 2.17: Framework for integrating data warehouses (Bruckner et al. 2001).

Pedersen et al. (2002) presented an approach that intends to match datacube content (represented in XML) with data provided by users (also represented in XML). For that, the authors introduced a federated query language, SQL_{XM} , incorporating the XML query language XPath¹². The authors avoided the semantic conflicts of heterogeneous datacubes.

Hümmer et al. (2003) also proposed an XML-based approach called XCube. XCube is an open, manufacturer independent and XML based family of document templates to store, query datacube, and exchange data between different datacubes. The approach introduced three XML documents for datacube elements: XCubeSchema, XCubeDimension and XCubeFact. XCubeSchema holds the multi-dimensional schema, XCubeDimension holds the hierarchical structure of the dimensions, and XCubeFact contains the fact data. The approach does not deal with semantic conflicts of heterogeneous datacubes.

Frank and Chen (2005) defined an approach that uses XML to store data derived from different data warehouses and uses XML Schema to define a global schema (i.e., an XML document generated from heterogeneous local datacubes). The proposed approach enables the integration of XML documents using XQuery which is an XML query language (Chamberlin 2002). Specifically, the approach consists of four steps:

1. predefine a global datacube schema by integrating local datacubes schemas,
2. transform each local datacube into an XML document,
3. manipulate the XML documents by pre-defined XQuery commands, and
4. integrate XML documents into a unified document (i.e., global datacube).

The proposed approach enables users to browse and interrogate a global datacube schema. However, the authors provided only some examples of how to solve semantic heterogeneities of datacube elements (e.g., examples of naming conflicts). Indeed, they focused on solving technical heterogeneity and paid less attention in dealing with semantic conflicts of heterogeneous datacubes.

¹² <http://www.w3.org/TR/xpath/>

Dubé et al. (2009) proposed a new XML grammar to facilitate the exchange of geospatial datacubes. It enables the delivery of the cube schema, dimension members (including the geometry of spatial members) and fact data. The use of this XML grammar has been demonstrated in the context of a Web Service. Also, this work does not address semantic conflicts of heterogeneous datacubes.

Representing datacube elements in XML format is very useful because data can be easily transferred from one datacube to another. However, the previous approaches are limited to solving structural heterogeneity, and avoided dealing with semantic conflicts of heterogeneous datacubes.

2.5.4 Perspective of existing approaches of interoperability

In order to deal with the heterogeneity problems, many approaches of interoperability between traditional databases, between geospatial databases and few between non-spatial datacubes have been developed (Brodie 1992, Goodchild et al. 1999, Bishr 1998, Harvey et al. 1999, Brodeur 2004, Staub et al. 2008) and many standards have been specified (GML, ISO/TC 211 19115:2003 Geographic information - Metadata). However, no work on interoperability between geospatial datacubes has been found in the literature.

The approaches for transactional databases focus on transactional aspect of data (i.e., normalised, detailed, application oriented data). While such approaches can be used to a certain extent (e.g., using a geographic ontology matcher to define the semantic similarity between the concepts of geospatial datacubes), they fell short in satisfying the interoperability between geospatial datacubes. In fact, first the elements of interest to be compared when interoperating datacubes are different from those of transactional databases. First, while the semantic interoperability of transactional databases deals with tables attributes and values, the semantic interoperability of datacubes deals with dimensions, hierarchies, levels, members, measures and facts). Second, unlike the semantic interoperability of transactional databases, the semantic interoperability of datacubes particularly stresses the importance of dealing with the semantics of aggregation and generalization relationships, the semantics of summarizing methods and algorithms, the semantics of summarizability conditions, the semantics of cross-tabulations for every level

of details and every member of the datacube dimensions and the semantics of geospatial hyper-cell which describes a model for a number of facts. Third, the semantic interoperability of datacubes deals with the context information related to decisional data content. Such context information may contain complex elements such as multidimensional integrity constraints which are assertions typically defined in order to prevent the insertion of incorrect data into datacubes.

For the interoperability of non-spatial datacubes, few works have been proposed (Bruckner et al. 2001, Mangisengi et al. 2001, Pedersen et al. 2002, Hümmer et al. 2003, Frank and Chen 2005). These approaches have been focusing on structural and technical aspects rather than the semantic conflicts of datacubes content.

In addition, while existing approaches of interoperability have proposed many solutions to deal with the semantic heterogeneity, they have adopted a non-systematic way (e.g., defining a semantic agreement between all concepts without analyzing its feasibility). As a result, stakeholders in the interoperability process (i.e., end-users, human mediator, or software agents) may fail to evaluate the causes and consequences of semantic problems, and to make decisions about responding to them (e.g., decide whether to solve the semantic problems, or to ignore them and endure potential consequences?). Consequently, existing approaches of interoperability still remain vulnerable to the semantic problems. In the interoperability between geospatial datacubes such problems are even more significant since they may cause faulty analysis or bad decisions (cf., chapter 4). So, it is very important to deal with the semantic problems in a systematic manner.

We should also notice that existing approaches of interoperability are usually based on ontology. Ontology-based approaches use Semantic Web technologies (e.g., RDF(S) and OWL) to represent and reason about data. These technologies are the most commonly used thanks to their formal structure and high expressiveness. However, ontology-based approaches focus on verifying the internal quality of the data and of context information (i.e., the extent to which data and context information is free from errors and inconsistency). However, in semantic interoperability, we need also to take into account the appropriateness of data and its context with regard to the application for which the interoperability is carried out. That is, the fitness-for-use of data and of its context.

The approach we propose in this thesis is based on evaluating the fitness-for-use of context information to help identifying the semantic problems and to help agents to make appropriate decisions with regard to such problems.

2.6 Conclusions

In this chapter, we presented and synthesized several topics related to the context of our research. Also, we reviewed key approaches of semantic interoperability between traditional databases, between geospatial databases and between datacubes.

We first reviewed the fundamental concepts of datacubes and their structures, and studied the difference between transactional databases and decisional datacubes. Then, we presented the importance of geospatial datacubes as an efficient component of DSSs.

Then, we presented the basic notions of interoperability (e.g., human communication, ontology, context, standards, geospatial data quality and semantic similarity), and we discussed the semantic heterogeneity problems that may occur during the interoperability.

After that, we reviewed key research on the semantic interoperability of different types of databases and studies their contributions and limitations. Key research on semantic interoperability was grouped according to the type of databases it is defined for, i.e., traditional and geospatial databases, as well as non-spatial datacubes.

Our literature review revealed that while existing approaches of semantic interoperability between transactional databases can be used to a certain extent, they present some limits with regards to the interoperability between geospatial datacubes. In fact, existing approaches do not stress the importance of dealing with the semantics of aggregation (e.g., semantics of methods, algorithms and conditions) and the semantics of cross-tabulations for datacube levels. Also, existing approaches deal with the semantic problems in a non-systematic manner that leaves them vulnerable to those problems.

For non-spatial datacubes, we saw that the few approaches attempting to support the semantic interoperability have focused on technical and structural aspects of interoperability. They are based on defining a common schema, usually, using XML, to

allow users to interrogate multiple datacubes. But, semantic interoperability requires more than describing the datacube schema in XML. It requires means to resolve semantic differences that lie behind apparently consensual representations.

Also, we reviewed the notions of context and context quality in semantic interoperability process. Any failure in providing data with acceptable context quality may undermine the reuse of decisional geospatial data i.e., the main aim of interoperability between geospatial datacubes. Consequently, we should take into account not only the context information, but also the quality of context in order to support the interpretation during the interoperability process involving different geospatial datacubes. However, no work on the quality of context with regards to semantic interoperability has been found. In this thesis, we define an approach to evaluate the quality of context associated with the content of datacubes.

Based on the limits of existing approaches in dealing with the semantic interoperability between geospatial datacubes and their components, new approach to support such interoperability is required. This approach should take into account the perspective of geospatial data used for strategic decisions. First, a question may be raised: do we really need to interoperate geospatial datacubes? We will suggest an answer to that question in the next chapter.

Chapter 3: Defining semantic interoperability between geospatial datacubes

3.1 Introduction

In the previous chapter we reviewed and synthesized several fundamental concepts related to geospatial datacubes and interoperability, and we reviewed existing approaches of interoperability between transactional databases and between non-spatial datacubes, with a special focus on the semantic aspect of interoperability. We saw that geospatial datacubes are considered efficient components of decision support systems. They enable making strategic decisions by supporting the user's mental model of data. They allow users to navigate aggregated and summarized data according to a set of dimensions with different hierarchies (Codd et al. 1993, Malinowski and Zimányi 2004, Rivest et al. 2005, Bédard and Han 2008). Geospatial datacubes contain geospatial, temporal and thematic data which may differ in format and content (e.g., geospatial characteristics such as location and geometry, and levels of abstraction) resulting in heterogeneous geospatial datacubes. The heterogeneity of geospatial datacubes presents an obstacle when people need to discover spatio-temporal trends or make strategic decisions using information located in different geospatial datacubes.

Interoperability has been widely recognized as an efficient paradigm for joining heterogeneous systems to facilitate an efficient exchange of information (Bishr 1998, Goodchild et al. 1999, Harvey et al. 1999, Brodeur 2004, Sboui et al. 2007). It aims at resolving technical, structural, and semantic heterogeneities between various systems in many fields (e.g., information management and engineering technologies).

In the last chapter, we noticed that existing approaches of interoperability still remain vulnerable to semantic problems. Moreover, although the interoperability of information systems and more especially in the geographic information realm has attracted the attention of many researchers (Brodie 1992, Bishr 1998, Harvey et al. 1999, Brodeur 2004), there have not been many works on the interoperability between non-spatial datacubes. Also, no

work on interoperability between geospatial datacubes has been found in the literature. The purpose of this chapter is to define the interoperability between geospatial datacubes, with a special focus on semantic aspects. More specifically, this chapter explains the need for interoperating geospatial datacubes (cf., section 3.2), defines key aspects of such interoperability (cf., section 3.3) and proposes a categorization of semantic problems that occur when interoperating geospatial datacubes (cf., section 3.4).

3.2 Need for interoperating geospatial datacubes

Geospatial datacubes are typically built for specific purpose. However, it may happen that we need to reuse the content of these datacubes for different purposes. For instance, in order to analyze the risk of forest fire behaviour and effects on the border region between *Canada* and *USA*, we would need to reuse different datacubes; for the purpose of this thesis, one was developed to analyze the risk of forest fire in south-east of Canada, and another that was developed to analyze the risk of forest fire in the north-east of United States. Discovering data trends and making strategic decisions would not have been as apparent by simply navigating separately through such geospatial datacubes as using interoperability which is widely recognized as an efficient paradigm for simultaneously re-using data from several scattered data sources.

Interoperability between geospatial datacubes may be required in many situations. We group these situations into three categories:

1. *Simultaneous and rapid navigation through different datacubes*: Users from different disciplines 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 and, hence, can considerably facilitate the navigation task. For example, the interoperability between geospatial datacubes would enable a common multidimensional model (i.e., without data records) that is

connected to the different dimensions of geospatial datacubes, or in other words an on-the-fly constellation-like structure relying on similar or identical dimensions. Such model is useful especially in emergency situations when users need to rapidly navigate through data stored in geospatial datacubes without preoccupying themselves too much with the problems of heterogeneity. An example of a situation is a natural disaster that affects adjacent jurisdictions. In such a situation, we may need to navigate through different geospatial datacubes, developed in these jurisdictions, in order to get the right information and act quickly at different levels (e.g., local, provincial, and federal) or at different domains (e.g., geographic and political). In such cases, interoperability between geospatial datacubes is crucial to help preventing catastrophic losses.

2. *Rapid insertion of data in a datacube*: While data in datacubes are usually collected from legacy systems, they can be imported from other heterogeneous datacubes (Bédard and Han 2008). We may need to rapidly insert new data (e.g., measures, members and member properties) in a geospatial datacube from other datacubes. An example of inserting measures would be to add the *electoral numbers* measure to a datacube *national election* from a datacube *local election*. An example of inserting members would be to add the lakes “Lac Saint-Jean” in a geospatial datacube about the construction of winter bridges from another that contains data about lakes. An example of inserting member properties would be to add the area of lakes and the area of roads in a geospatial datacube about the construction of winter bridges, from two other geospatial datacubes; one of them contains data about traffic and the other contains data about lakes.
3. *Interactive and rapid comparison of scattered decisional data to analyze phenomena changes*: In order to analyze phenomena change (e.g., forest stand dynamics), we need to compare data describing these phenomena at different epochs. We may need to compare data stored in geospatial datacubes built also at different epochs. Interoperating geospatial datacubes would permit interactively comparing data and analyzing changes. For example, in order to analyze changes in wood volume following a natural disaster, we may need to rapidly compare

information about forest stands stored in different geospatial datacube. Interoperability enables rapid navigation through these datacubes and detection of changes in the volume of wood.

However, data in datacubes are usually collected from different source systems or from different versions of a same data source (e.g., System A as existing in 2003 and System A as existing now). One may ask “why shouldn’t we just interoperate the transactional source systems rather than datacubes and use existing interoperability solutions?” There are three main reasons for interoperating geospatial datacubes:

1. We possibly no longer have access to data source systems from which we created the datacubes due to multiple reasons including administration policies (e.g., backup practices, security), the replacement of past systems by new systems (software and/or hardware), systems merging after company’s reorganisation (e.g., municipal mergers in Canada during the early 2000s), retirement or departure of key employees, and the non availability of data sources (e.g., bankrupt companies, data destroyed).
2. We need to use data from a long period (i.e., historic data) that usually exist only in datacubes. In fact, in source systems, past data are usually modified or replaced by new data and then destroyed or archived, whereas datacubes keep historic data for strategic decision-making purposes (Bédard and Han 2008). So if we need to reuse such data, we have sometimes no choice but to consider datacubes.
3. In the context of decision-making, interoperating geospatial datacubes is potentially more efficient than interoperating source systems. In fact, within a geospatial datacube, contrary to source systems, possible aggregations of measures for all possible combinations of members are pre-calculated using different operators (e.g., mathematical, metric such as distance and area, and topological relations such as disjoint and interior intersection) (Rivest et al. 2005). These aggregations usually require an arduous work for geospatial datacubes developers (e.g., defining procedures for aggregation, defining a new

geospatial level as an aggregation of others). Re-using datacubes means that we don't need to redefine such aggregations from scratch, thus saving time and money.

3.3 A new definition of interoperability between geospatial datacubes

As seen in the previous chapter, geospatial datacubes aim at supporting a high-level decision-making process. They enable users to navigate within the different levels of granularity of geospatial data. A geospatial datacube has three main levels of abstraction: 1) *cube* includes cubes, measures and dimensions, 2) *dimension* includes hierarchies and levels of dimensions, and 3) *member* includes members of dimensions.

With these characteristics in mind, and inspired especially by the definition of interoperability provided by Brodie (1992), **we define the interoperability between two geospatial datacubes C1 and C2 as the ability of C1 to request a service in a manner that can be understood by C2, and the ability of C2 to respond to that request in a manner that can be understood by C1 and adapted to its context.** The request and response are conducted automatically. Services could include:

- importing/exporting geospatial members contained in a datacube element (i.e., cube, measure, dimension, or level);
- participating in the creation of a new geospatial datacube element (e.g., dimension and level);
- getting information about a geospatial datacube element (e.g., the type of method used for a geospatial measure);
- verifying the change of a geospatial datacube element (e.g., change of definition, of a geometric representation).

These services involve one or more of the following categories of actions:

1. *Category 1*: Comparing an element of a geospatial datacube (i.e., measure, dimension, hierarchy, or level) against an element of another geospatial datacube. For example, comparing the dimension *Territory* against the dimension *Administrative region*, or comparing the measures *population density* and *number of people*. In this case, the interoperability between datacubes would lead to a comparison report (e.g., a comparison report of the changes in population density in different areas).

2. *Category 2*: Updating an element of a geospatial datacube based on the content of other datacubes involved in the interoperability process. This may include modifying one or more semantic aspects of the datacube element (e.g., modifying the definition of a measure, changing the context of a dimension). It may also include updating geospatial members, updating member properties, or updating the values of some measures. Updating a datacube element during the interoperability process may involve the modification of other datacube elements. For example, slightly changing the definition of the dimension *Road* to exclude railroads, may involve the modification of the measure *Number of accidents*. In this category, the interoperability between datacubes would lead to the update of one or more datacubes involved in the interoperability process.

3. *Category 3*: Integrating datacubes involved in the interoperability process. In this case, the interoperability between datacubes would lead to the creation of one or more federated geospatial datacubes. This refers basically to the creation of a common conceptual model from the models of the datacubes involved in the interoperability process. Such a common model would allow access and virtual navigation through existing datacubes as if they were only one, or creating a new datacube that will contain data extracted from existing datacubes. This category may involve the following actions:
 - 3.1. *Integrating measures*. Integrating measures may refer to one of the following two actions:
 - a. The first consists of adding a new measure to a datacube from another one based on common dimensions and members. For example, adding in a datacube *Local*

election a measure *Vote-total* from another datacube *National election*, see Figure 3.1. In order to perform such integration, we need to slice the dimensions specific to the datacube source to obtain only the measures according to the common dimensions of both the datacube source and the datacube destination. For example, slicing the dimension *Time* of the *National election* datacube to obtain only the electoral results according to age category and region.

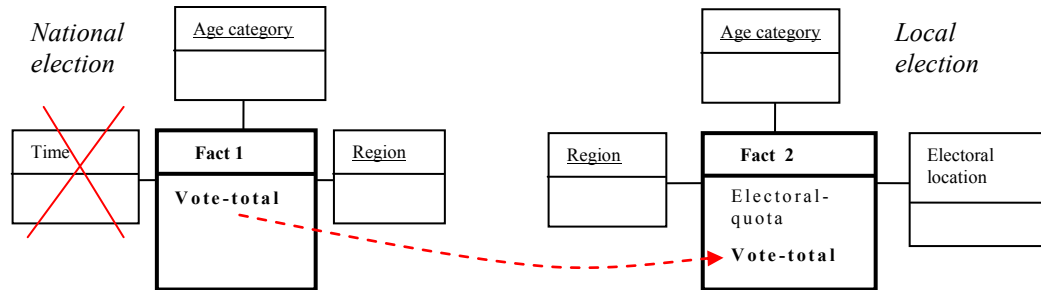


Figure 3.1: Inserting a measure from another datacube.

- b. The second consists of creating a new common measure by merging existing measures of different geospatial datacubes involved in the interoperability process, or by choosing one (or more) measures among existing semantically related measures. An example of this action is the creation of the measure *population density* based on the two measures *Area* of datacube *C* and *number of people* of datacube *D*. Also, creating the geometric measure *Zone & intersection*, with a complex geometric primitive, based on the two geometric measures *Intersection point* of datacube *Intersection* and *Zone accident* of datacube *Accident* (see Figure 3.2).

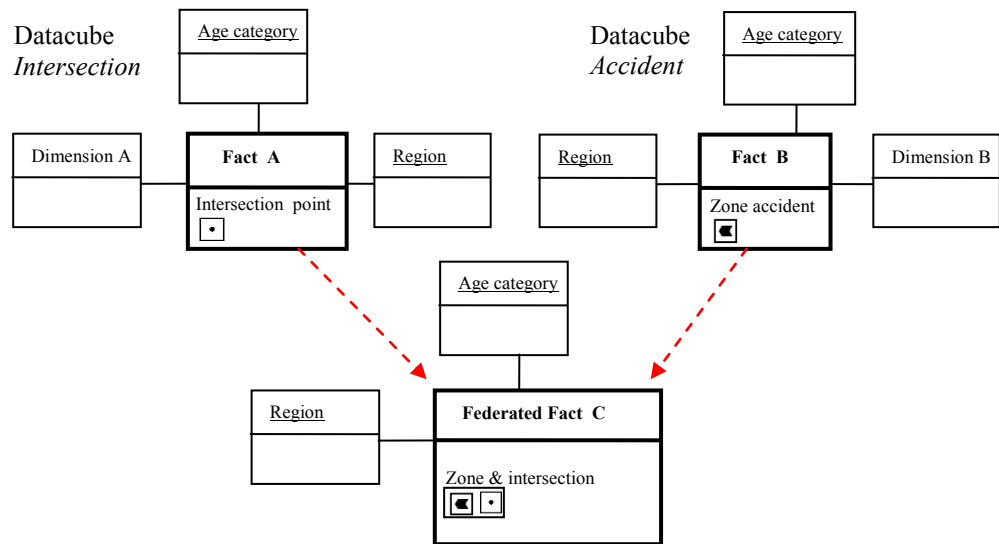


Figure 3.2: Creating a new measure from two existing measures.

An example of choosing one among existing semantically related measures is to select the measure *Intersection point* from the measures: *Intersection point* and *Zone accident* (see Figure 3.3).

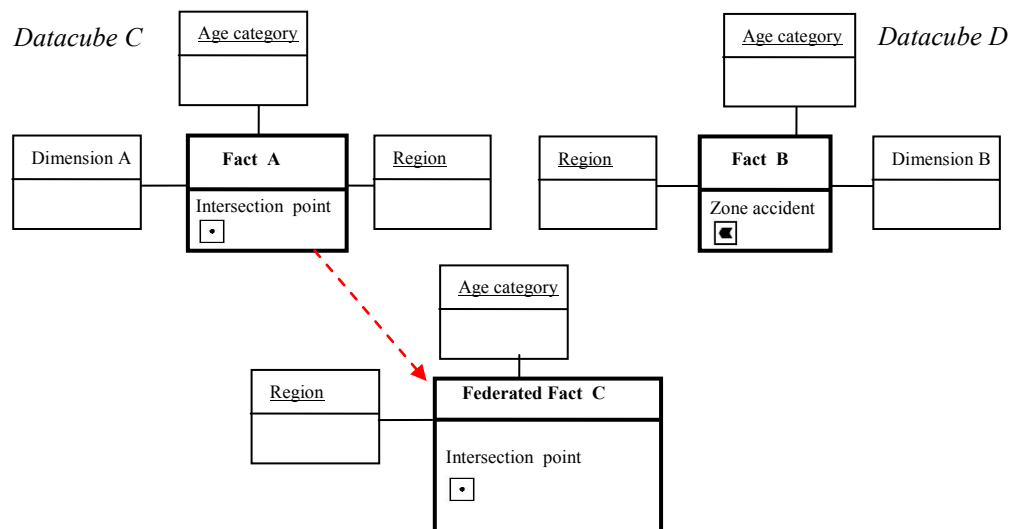


Figure 3.3: Choosing one of two existing measures.

3.2. *Integrating dimensions*. Integrating dimensions may refer to one of the following three actions:

- a. The first action consists of adding one or several dimensions of one datacube to another. Since, in a datacube, the values of measures depend on the context set up by the dimensions, adding a new dimension to a datacube involves updating the measures that depend on that dimension. For example, adding to the *Human resources* datacube a dimension *Age category* from the *Manpower* datacube involves recalculating the number of employees taking into account the members of this dimension (see Figure 3.4).

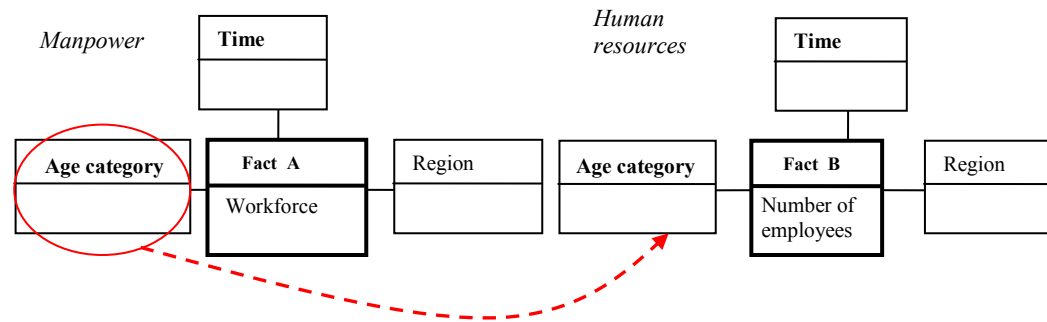


Figure 3.4: Adding a dimension to datacube from another.

- b. The second consists of creating a new dimension by merging existing dimensions of different datacubes, or by choosing one or more dimensions among existing semantically related dimensions. An example of dimension merging is the creation of a dimension *Province-State* for the new datacube *USA-Canada* by combining the dimensions *Province* and *State* of the datacubes *Canada* and *USA* respectively, see Figure 3.5. The resulting dimension will contain members of both dimensions (*Province* and *State*).

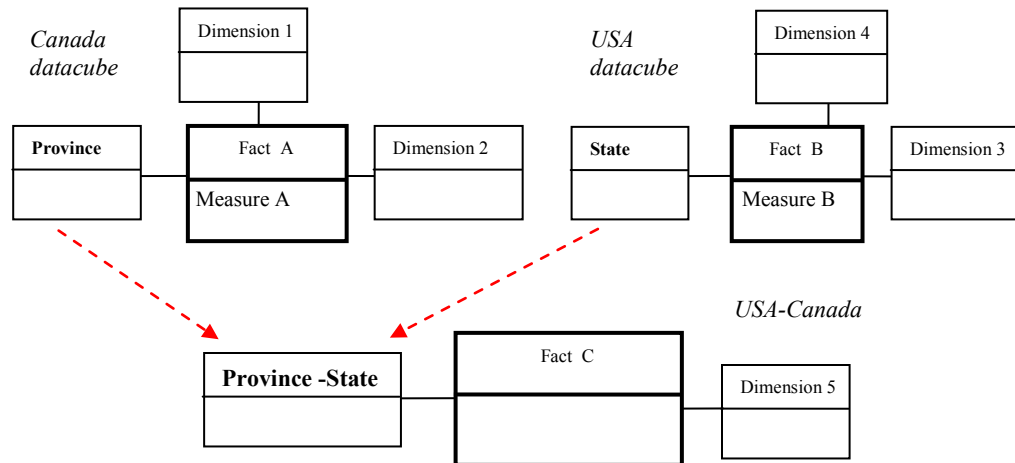


Figure 3.5: Creating a new dimension from two existing dimensions.

An example of choosing one among existing semantically related dimensions is to select the dimension *Residential zoning* from the dimensions: *Residential zoning* of the datacube *Quebec 1982* and *Municipal zoning* of the datacube *Quebec 2000* (see Figure 3.6).

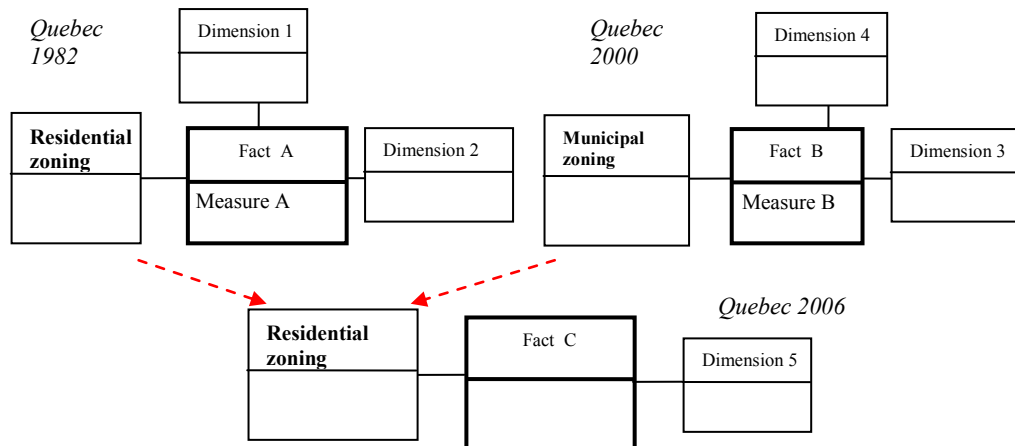


Figure 3.6: Choosing one of two existing dimensions.

As for the previous case, creating a new dimension based on existing dimensions of different datacubes involves updating the measures that depend on that dimension.

- c. The third action consists of modifying a dimension of a datacube by using one or more of the dimension's levels of another datacube. For example, adding the level *State* to transform a simple hierarchy *Territory* for one country to a multiple hierarchy that takes into account the characteristics of territorial divisions of each country in the datacube *USA-Canada*, as shown in Figure 3.7.

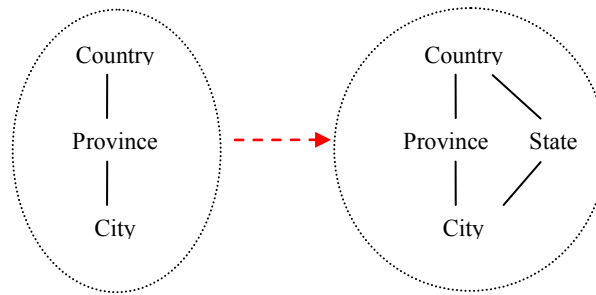


Figure 3.7: Transformation of a simple hierarchy *Territory* for one country to a multiple hierarchy by adding the level *State*.

We should notice that performing one or more of the previous actions may bring about other actions. For example, after adding a new dimension to a datacube, we may need to add new values of the measures of this datacube (e.g., adding the dimension *region* to a datacube that contains information about road accidents involves adding new values to the measure *number of accidents* according to the dimension *region*).

We also would like to point out the relationship between the notions of interoperability and integration of geospatial datacubes. In fact, as seen in the third category of actions (i.e., integrating datacubes involved in the interoperability process), interoperability between geospatial datacubes can serve to integrate these datacubes. In this case, **integration is an outcome of the interoperability process**. Moreover, one can reuse such an outcome in further interoperability between these datacubes. In this case, **integration becomes a way to assist the interoperability process**. This shows a key relationship between interoperability and integration of geospatial datacubes in particular and of databases in general; interoperability between datacubes can be performed to achieve an integration of these datacubes, while, the integration of datacubes can be used to facilitate the interoperability process between them.

Whatever the outcome of interoperability between datacubes is (i.e., integrated datacubes, update of one or more datacubes, or a comparison report of different phenomena), it should take into account the *raison d'être* of the multidimensional paradigm which is the ability to provide a strategic decision support (i.e., providing users with relevant elements to make strategic decisions). In other words, any result of such interoperability should support the user's mental model of data and allows him/her to rapidly navigate through different levels of detail.

As it is the case for the interoperability between transactional databases, interoperability between geospatial datacubes must deal with semantic problems. Categorizing such problems would facilitate dealing with them.

3.4 Semantic heterogeneity in geospatial datacubes

In previous chapters, we have seen that semantic heterogeneity represents a major challenge for enabling interoperability between information systems. This heterogeneity is caused basically by differences in data description, or the lack or the inappropriateness of context information. Semantic heterogeneity appears more significant when dealing with geospatial datacubes. This due to the fact that such heterogeneity includes, in addition to traditional non-spatial differences, 1) differences in geospatial characteristics (e.g., geometry, graphics), and 2) differences in multidimensional characteristics (e.g., dimensions with different levels, different aggregation methods, and different details of complex ETL procedures). In order to support the interoperability between geospatial datacubes, we need to classify the problems related to semantic heterogeneity that may occur in them.

As in transactional databases, semantic heterogeneity in geospatial datacubes occurs when there are differences in **schema elements and in production context information (metadata)**. While schema heterogeneity refers to the difference in structure of datacube elements (e.g., hierarchy structure, number of levels), context heterogeneity refers to the difference of all other elements of the datacube model. Examples of schema heterogeneity include the difference in the geometric primitives used to represent the members of two semantically related levels, and the difference in the number of levels of two semantically

related dimensions. Examples of context heterogeneity include the difference in the methods used for data aggregation, and the difference in referencing systems.

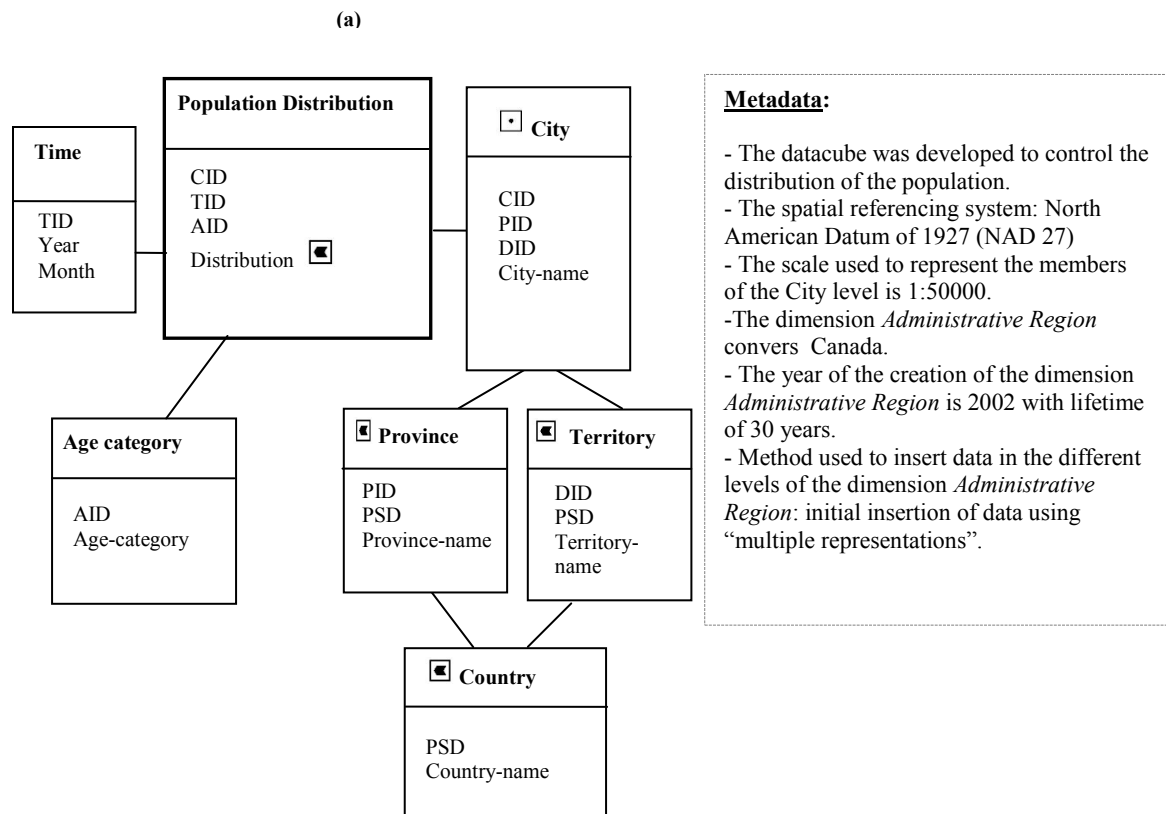
To illustrate our categorization, we present an example of using two geospatial datacubes (see Figure 3.8). To determine the risk of forest fire to a population, we can interoperate two geospatial datacubes C1 and C2 (modelled respectively in Figure 3.8 (a) and Figure 3.8 (b)). The first datacube C1 is used to analyze the distribution of the population in specific areas and periods. The second geospatial datacube C2 is used to analyze 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.

C1 contains three dimensions (*Time*, *Age category*, and *Administrative region*) and a fact table (*Population Distribution*) with a geometric geospatial measure (*Distribution*), which indicates clusters of people in geographic areas, and foreign keys related to the different dimensions. The dimension *Time* has a hierarchy with two levels: *Month*, and *Year*. The dimension *Age category* has a hierarchy with only one level titled *Age category*. Finally, the dimension *Administrative region* has a hierarchy which contains three levels ordered as follows: *City*, *Province*, *Territory*, and *Country*. We should notice here that a hierarchy represents an analytic perspective within a dimension, as perceived by Salehi (2009). Accordingly, labelling a set of ordered levels as a hierarchy depends on the datacube's analysis requirements. In this example, we need to analyze Canadian territories and Canadian provinces similarly. Hence, all the levels (*City*, *Province*, *Territory*, and *Country*) that are related to our analysis belong to the one hierarchy. However, if one needs to analyze Canadian territories and Canadian provinces separately, he/she would consider two hierarchies for these four levels: (1) *City*, *Province*, and *Country*, and 2) *City*, *Territory*, and *Country*.

C2 contains three dimensions (*Period*, *Region* and *Forest stand*) and a fact table (*Forest Fire*) which allows to analyze the regions vulnerable to the fire. This table contains a geometric geospatial measure titled *Fire zone* which indicates the zones that are vulnerable to forest fire. The dimension *Period* has a hierarchy with two levels: *Month*, and *Year*. The dimension *Region* has two hierarchies: a hierarchy with four levels: *City*, *Province*, *Territory* and *Country*, and another hierarchy with three levels: *City*, *State* and *Country*.

The dimensions *Forest stand* has one hierarchy which contains only one level (*Group*). The geometric geospatial levels as well as geometric geospatial measures are modelled using geospatial pictograms. A pictogram is a symbol that refers to geometric primitives in geospatial data modeling (Bédard and Larrivée, 2008). In our example, we use the pictograms developed in Perceptory¹³ tool where the pictogram «□» represents a 0D type, the pictogram «▣» represents a 1D type, and the pictogram «■» represents a 2D type. For example, in the model of Figure 3.8 (a), the geometries of the levels of *Administrative region* (*City, Province, Territory and Country*) are modelled using respectively one pictogram «□» and three pictograms «■». The geospatial measure *Distribution* is modelled using the pictogram «■».

In our example, the production context of the geospatial datacubes contain information related to the geospatial referencing system, scale and precision, year of creation, method of forest stand measurement as well as the geospatial coverage of some dimensions.



¹³ Perceptory's Web Site: <http://sirs.scg.ulaval.ca/perceptory>

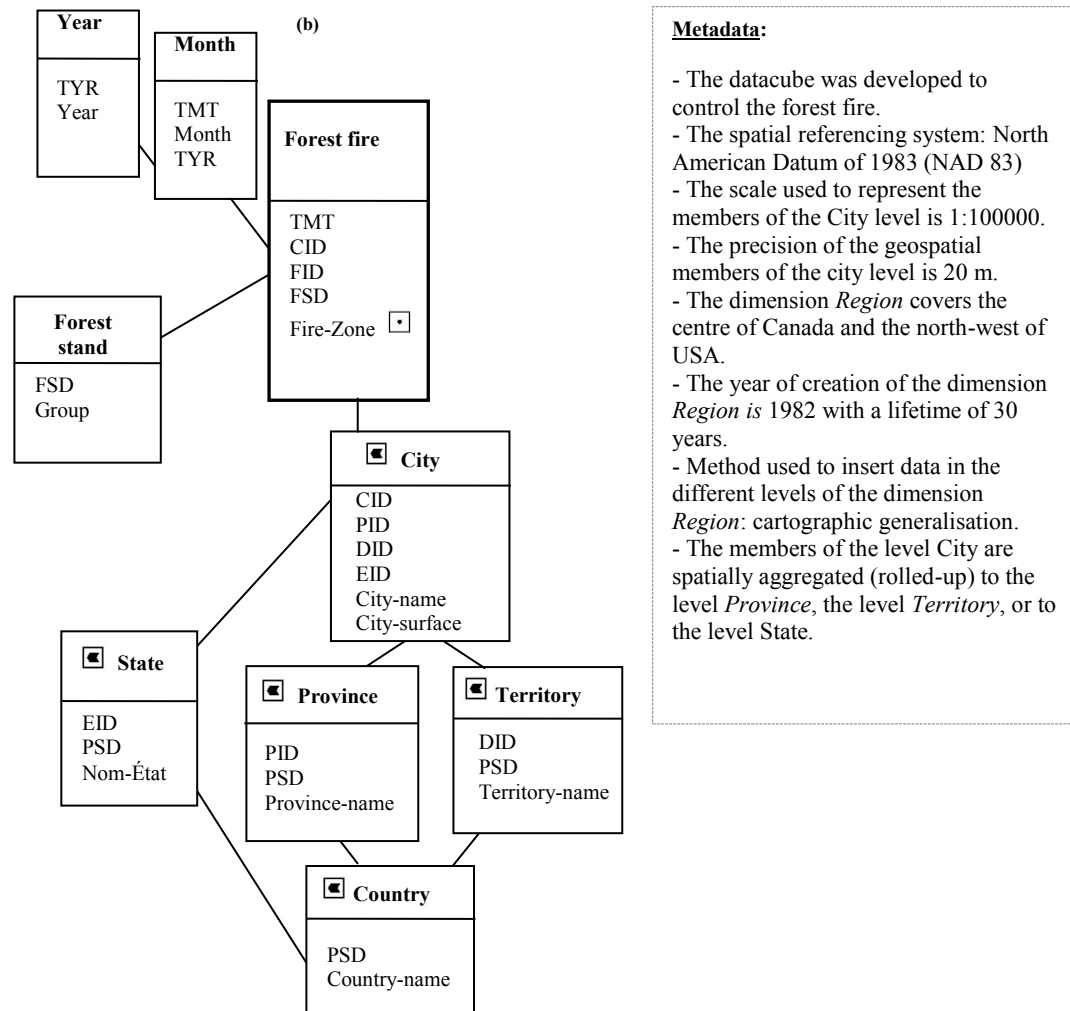



Figure 3.8: Two models of datacubes (C1 and C2).

The problems of semantic heterogeneity of the models of geospatial datacubes can exist at five different levels: cubes, measures, dimensions, hierarchy and levels. The latter level of heterogeneity involves the semantic difference in geospatial members. We should notice here, that as we point out in chapter 2, we believe that the geometric and graphic aspects of members (e.g., the form and the orientation) may convey semantics when associated with an ontological definition (e.g., a legend). For example, according to its definition within an ontology, a cartographic element may refer to houses according to their roofing or to houses according to their foundation (see Figure 2.8). Therefore, the geometric and graphical aspects of members may be considered as part of semantic heterogeneity.

However, based on the literature review, it was clear that the semantic problem of interoperability is very broad, and no single thesis could propose a solution that could take into account all aspects of semantics and achieve a perfect interoperability. Thus, we decided to focus on the “ontological aspects” of the geospatial datacube members as their semantic vehicle. The “ontological aspects” are generally associated with the semantics of member’s level,

In order to categorize the problems of heterogeneity, we formalize the elements of datacubes. Let consider C, D, H, L, and A, respectively, a datacube, dimension, hierarchy, level and an attribute.

1. An attribute a is defined by the triplet $a = (primitive, nature, domain)$ where:
 - $primitive$ is a textual term (e.g., “surface”), a numeric, a geometric (e.g., point, line and polygon), or a date (e.g., instant and interval).
 - $nature$ refers to the spatial, temporal, or thematic nature of the attribute.
 - $domain$ is the domain of attribute’s values.

The nature of the attribute indicates “what” the attribute represents. The primitive indicates “how” the attribute is represented. For example the spatial attribute fire zone is represented using a geometric primitive  and the textual term “Fire_zone”.

2. A level $L = \{a_1, a_2, \dots, a_n\}$ is a set of n attributes where n is the cardinality of the set. For example, the level $City = \{city_name, city_surface\}$.
3. A hierarchy $H = (\{L_1, L_2, \dots, L_h\}, <)$ is a set of h levels and an order relation between these levels. h is 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* $= (City < Province < Country)$.

4. A dimension $D = \{H_1, H_2, \dots, H_d\}$ is a set of d hierarchy. d is the cardinality of the set and is in most cases equal to 1. For example, the cardinality of the dimension *Establishment* = $\{(Unit < Group\ of\ units < Department)\}$ is 1.
5. A measure m is an attribute which describes the subject of analysis. It is the dependent variable that depends upon the independent variables (the members in D).
6. A datacube $C = (\{D_1, D_2, \dots, D_c\}, \{m_1, m_2, \dots, m_q\})$ is a set of c dimensions (with their members) and q measures.

We distinguish five categories of heterogeneity that may occur between the models of geospatial datacubes. For each category, heterogeneity can occur in schema and in production context information (metadata). We should note that, identifying the heterogeneity problems of production context can be extremely difficult because context can include various aspects (e.g., methods of aggregation, data precision and measuring techniques) which are usually not mentioned explicitly. Only some examples are mentioned in the following categorization.

1. Cube-to-Cube heterogeneity

- a. *Schema Heterogeneity:*

- It appears when two datacubes have semantically related measures which are defined according to various numbers of dimensions. For example, the accident frequency can be analyzed according to three dimensions (*Road, Region, and Time*) in $C1$, and to four dimensions (*Road, Region, Type of vehicle and Time*) in $C2$.

- b. *Context Heterogeneity:*

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

- *Difference of data source systems.* It appears when two datacubes dealing with semantically related subjects of analysis were created from two different data source systems.

2. Measure-to-Measure heterogeneity

a. Schema Heterogeneity:

- *Difference in type.* It appears when two semantically related measures are represented according to different types (e.g., nominal, numeric or geometric).
- *Difference in geometric primitives.* It appears when two semantically related measures have various geometric primitives (e.g., point vs. polygon).

b. Context Heterogeneity:

- *Heterogeneity of calculation function.* Appears when different functions were used to calculate semantically related geospatial measures. For example, the function geometric union is used to calculate the measure *density* in a datacube C1, whereas the function centre of gravity is used to calculate the measure *concentration* in datacube C2. *Density* and *concentration* are two semantically related measures.

3. Dimension-to-Dimension heterogeneity

a. Schema Heterogeneity:

- *Inequality in the number of hierarchies.* This appears when the cardinalities of semantically related dimensions are different. More precisely, let us suppose that n_1 is the cardinality of dimension D1, n_2 is the cardinality of dimension D2. If $n_1 \neq n_2$ then there is a heterogeneity Dimension-to-Dimension at the schema level. In our

example, the dimension *Administrative region* has only one hierarchy contains: (Country, Province, and City), whereas the dimension *Region* has two hierarchies contain: (Country, Province, Territory, and City) and (Country, State, and City).

b. Context Heterogeneity:

- *No-correspondence of the dimension constraints*¹⁴. This appears when constraints of two semantically related dimensions are incoherent. For example, the constraint of dimension *Administrative region* of the datacube C1 indicates that all the members of the level *City* are rolled-up to the level *Province*, while the constraint of dimension *Region* of the datacube C2 indicates that the members of the level *City* are rolled-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*. This appears when the cardinalities of the hierarchies of semantically related dimensions are unequal. More precisely, let n_1 be the cardinality of the hierarchy $H1$ and n_2 the cardinality of the hierarchy $H2$. If $n_1 \neq n_2$, then there is a schema heterogeneity (inequality of the number of levels). For example, 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 the dimension *Region* contains three levels.
- *Inequality of order of levels*. This 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, we may have different orders of the levels *Sex*, *Age-group*, and *All*: 1) *Sex* < *Age-group* < *All*, and 2) *Age-group* < *Sex* < *All*, though this is not common nor is perceived as good design practice. Also, 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*.

b. Context 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, Province, Territory and City) covers the center of Canada and the North-West of the United States.
- *Heterogeneity between the methods used for aggregating the members of different levels.* It appears when the members of the levels belonging to two semantically related hierarchies are aggregated using two different methods. For example, the members of the levels belonging to the hierarchy (Country, Province, Territory and City) of the dimension *Administrative region* are aggregated using multiple representation, whereas the members of the levels belonging to the hierarchy (Country, Province, Territory and City) of the dimension *Region* are aggregated using cartographic generalization.

5. Level-to-Level heterogeneity

¹⁴ Hurtado et al. (2005) introduced the notion of dimension constraint.

a. *Schema Heterogeneity:*

- *Inequality of the number of attributes.* It appears when the cardinalities of semantically related levels are different. More precisely, let n_1 be the cardinality of the level $N1$ and n_2 be the cardinality of the level $N2$. If $n_1 \neq n_2$ then there is a schema heterogeneity Level-to-Level. For example, the geospatial level *City* in C1 is described using one attribute: *City-name*, whereas the level in C2 is described using two attributes: *City-name* and *City-surface*.
- *Difference in geometric primitives.* It arises when, in two geospatial datacubes, two semantically related levels have different types of geometric primitives. For example, in datacube C1, each member of the level *City* is represented using a point, whereas in datacube C2, each member of the same level *City* is represented using a polygon.

b. *Context Heterogeneity:*

- *Difference in geospatial referencing system.* It occurs when there is a difference of geospatial referencing systems used to determine the position of the geospatial members of semantically related levels. In our example, the levels of the dimension *Administrative region* are based on the North American Datum 1927 (NAD 27) system, whereas those of the dimension *Region* are based on the North American Datum 1983 (NAD 83) system.
- *Heterogeneity of cartographic scale.* It appears when semantically related levels are originally represented with different cartographic scales. In our example, the cartographic scale of the level *Province* of datacube C1 is 1:50000, whereas the scale of the level *Province* of the datacube C2 is 1:100000. Although one may represent both

levels at the same scale, the resolution of the details won't be the same since C1 will be generalized (thus simplified) to be presented at the scale of C2, or C2 will typically show more details than C1 when scaled up to the same scale.

- *Heterogeneity of precision.* It appears when semantically related levels are represented using different precisions. For example, the levels *City* of two datacubes C1 and C2 may have different precisions (10 m in datacube C1 and 20 m in datacube C2).
- *Temporal heterogeneity.* It appears when members of semantically related levels are collected at different periods. For example, Figure 3.9 shows that municipalities in the same geospatial coverage are represented differently at different epochs; following the administrative fusion of the two cities *A* and *C* in 2006.

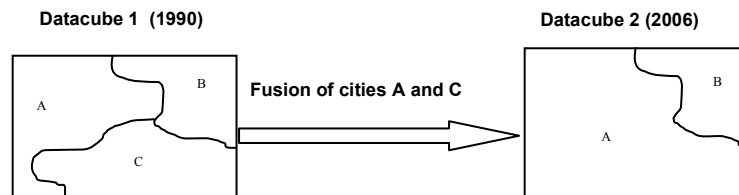


Figure 3.9: Temporal heterogeneity: fusion of municipalities *A* and *C* in 2006.

We remind that, in the geospatial domain, identifying the heterogeneity problems of production context can be extremely difficult because context can include potentially infinite number of aspects which are usually not mentioned explicitly. Only some examples were mentioned above but the aspects of context heterogeneity can include many others, such as geospatial positioning methods in referencing systems, techniques of data acquisition, methods of observations, algorithms of data pre-processing, algorithms of data transformation and data processing, cartographic generalization, and so on.

3.5 Conclusions

In this chapter, we explained the needs for interoperating geospatial datacubes. The situations where such needs arise are grouped into three categories: 1) a simultaneous and rapid navigation through different geospatial datacubes, 2) a rapid insertion of data in a datacube while, and 3) an interactive and rapid analysis of phenomena changes. Then, we defined key aspects of such semantic interoperability. In fact, interoperating geospatial datacubes may involve one or both of the following actions on their components: 1) comparing an element of a datacube (i.e., measure, dimension, hierarchy, and level) against an element of another datacube, 2) updating an element of a geospatial datacube based on the content of other datacubes involved in the interoperability process, or 3) integrating datacubes involved in the interoperability process. Finally, we proposed a categorization of semantic problems (i.e., semantic heterogeneity) that may occur in geospatial datacubes. The defined categories are: cube-to-cube heterogeneity, measure-to-measure heterogeneity, dimension-to-dimension heterogeneity, hierarchy-to-hierarchy heterogeneity, and level-to-level heterogeneity. In each category, we distinguished between schema heterogeneity and context (metadata) heterogeneity. This categorization will help us to define an approach to support the semantic interoperability between geospatial datacubes.

In the next chapter, we will propose a conceptual framework to support specifically the interoperability between geospatial datacubes.

Chapter 4: A conceptual framework for semantic interoperability between geospatial datacubes and associated risks

4.1 Introduction






In chapter 3, we motivated the need for interoperability between geospatial datacubes and we proposed a categorization of the semantic problems that may occur during this process. Moreover, in the chapter 2, we saw that existing approaches of database interoperability are not very suitable for the interoperability between geospatial datacubes. Consequently, in order to effectively support the interoperability between geospatial datacubes, we need a specific framework that deals with the specific characteristics of geospatial datacubes.

In this chapter, we introduce a specific framework for the semantic interoperability between geospatial datacubes based on human communication. The choice of the human communication model stems from our belief that people communicate efficiently. They interact using different representations of real-world phenomena, and usually end up understanding each other. Many authors recognized that human communication is an ideal model for the interoperability of information system (Bédard 1986, Brodeur 2004, Kuhn 2005). Nevertheless, people may need an interpreter (a mediator) to help them understanding the messages exchanged between them. Accordingly, in this chapter we propose a conceptual framework which is based on a mediator agent (context agent) to support the interoperability between geospatial datacubes. However, even with the intervention of a mediator, there are possible risks that may exist during the interoperability process between geospatial datacubes.

4.2. A conceptual framework for the semantic interoperability between geospatial datacubes

With the emergence of software agents, semantic interoperability has been viewed as the technical analogue to human communication (Brodeur 2004, Kuhn 2005). According to this

view, each receiver agent tries to interpret the exchanged data as it has been originally intended by the source agent, and then to adapt it to his/her/its specific context. We adopt this view and we define an agent communication framework which is based on software agents: *datacube agents* and a *context agent*. Datacube agents represent geospatial datacubes involved in the interoperability process, and engaged in a process of human-like communication. In order to facilitate the communication between datacube agents, the context agent mediates between them to help them interpreting the exchanged data, see Figure 4.1.

Let us assume the following situation that illustrates the interoperability between two agents representing geospatial datacubes. An agent representing the datacube C1 (A_{c1}) is looking to export the members of its level *Province* to the dimension of another datacube. He/she translates the name, the geometric representation of the level *Province* (“province” ) and the context associated with this level (spatial coverage: “Canada”) into a message (e.g., “province , Canada”) posted to an emergency control center which contains the datacube C2. An agent representing the datacube C2 (A_{c2}) receives and interprets the message, searches for related elements in his/her/its knowledge base. The search consists of comparing the semantics of the received data: “province ” with the semantics of each element of his/her/its knowledge base. If there is a semantic similarity between the received data with one or more elements of his/her/its knowledge base, A_{c2} asks a context agent (e.g., an emergency manager) to verify that the context associated with the received data (“Canada”) is relevant with the context associated with the result the A_{c2} has found. In this case, the context associated with the received data is relevant to the context associated with the result (“province” , “Canada”). Then, A_{c2} may send a confirmation to A_{c1} containing the following message: “There is only one similar element: province”. We should notice that, since the context agent confirmed that both contexts are similar, A_{c2} can go ahead and use the received message, and hence avoiding to send the notification to A_{c1} and probably wait for a confirmation, etc.

We believe that this way of communicating helps to enhance the semantic interoperability between geospatial datacubes. It uses a mediator (i.e., context agent) to make datacube agents aware of the context associated with the exchanged data between them. Being aware

of the context may save time and money. For example, knowing that the context of received data is relevant to his/her/its context, a destination agent may go ahead and use this data (without taking time to send a feedback to the source agent or to wait for his/her/its response). This way of interoperating may be even essential in some situations (e.g., when the source agent is not available anymore to provide more information to the destination agent). In such situations, the destination agent can rely on context agent to verify the relevance of the received data and its associated context to his/her/its specific need.

Context agent is responsible mainly for:

1. Comparing the concepts of both source context and destination context to determine if they “fit together”.
2. Making agents aware of the relevance of context information to their specific use.

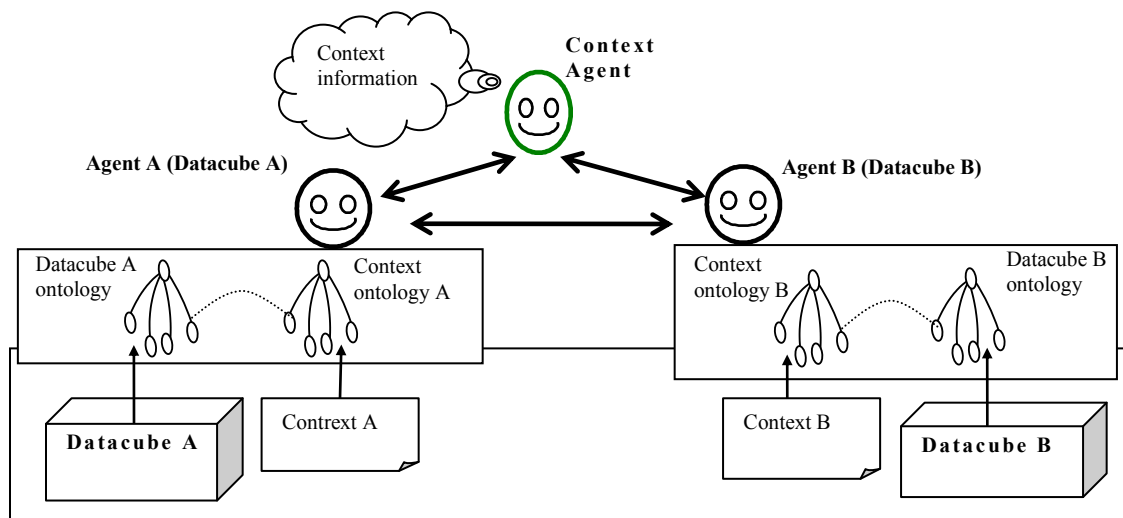


Figure 4.1: An overview of the communication between agents representing geospatial datacubes.

We define datacube agents according to five communication layers as shown in Figure 4.2. These layers correspond to 1) the five conceptual levels of geospatial datacube; from the more general level to the most detailed level (cube, measure, dimension, hierarchy, and

level), and to 2) the five categories of semantic problems that may occur during the interoperability between geospatial datacubes (cube-to-cube, measure-to-measure, dimension-to-dimension, hierarchy-to-hierarchy, and level-to-level). Each agent is responsible for resolving the semantic problems that may occur at one layer (e.g., the dimension agent is responsible for resolving the semantic problems that may occur at the dimension layer). As such, if the datacube is maintained by different parties (e.g., departments of an organization), it will be possible to trace the responsibility of each party in the interoperability process. For example, suppose that a party P_1 has to maintain a dimension d_1 of a datacube D_1 , and that P_1 can determine the behaviour of the agent representing the dimension d_1 . Then, P_1 will be responsible if, for instance, this agent withdraws from the interoperability process between D_1 and another datacube. We should note that the semantic problems 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.

The content of each geospatial datacube is described using two ontologies: one represents the concepts of geospatial datacubes and another represents the context related to the datacube's content. Based on these related ontologies, each agent representing one datacube's level can communicate with another agent representing the same level of another datacube (e.g., agent representing a dimension of a datacube A communicates with an agent representing the dimension of a datacube B). We call it *horizontal communication*. This communication is indicated in Figure 4.2 by solid bidirectional arrows.

In order to facilitate such communication, the context agent stores relevant information about exchanged data and use it to solve semantic conflicts that may exist between datacube agents (e.g., context mismatch). Also, a context agent has an ontology that describes generic context. This ontology is used basically to type the context of each datacube and to facilitate the match between different concepts of the context ontologies associated with heterogeneous datacubes content. The generic context has two other advantages: 1) defining the generic aspect of context information, and 2) guiding the categorization of the context associated with the content of each geospatial datacube. In

Annexe A, we propose an example of generic concepts of context (e.g., aggregation, referencing system and scale).

Besides the horizontal communication between agents representing the elements of heterogeneous geospatial datacubes, the framework defines a *vertical communication* between agents representing different levels within the same datacube (e.g., agent representing a dimension of a datacube A communicates with an agent representing a measure of the same datacube). In Figure 4.2, the dashed bidirectional arrows indicate the vertical communication between agents within the same datacube.

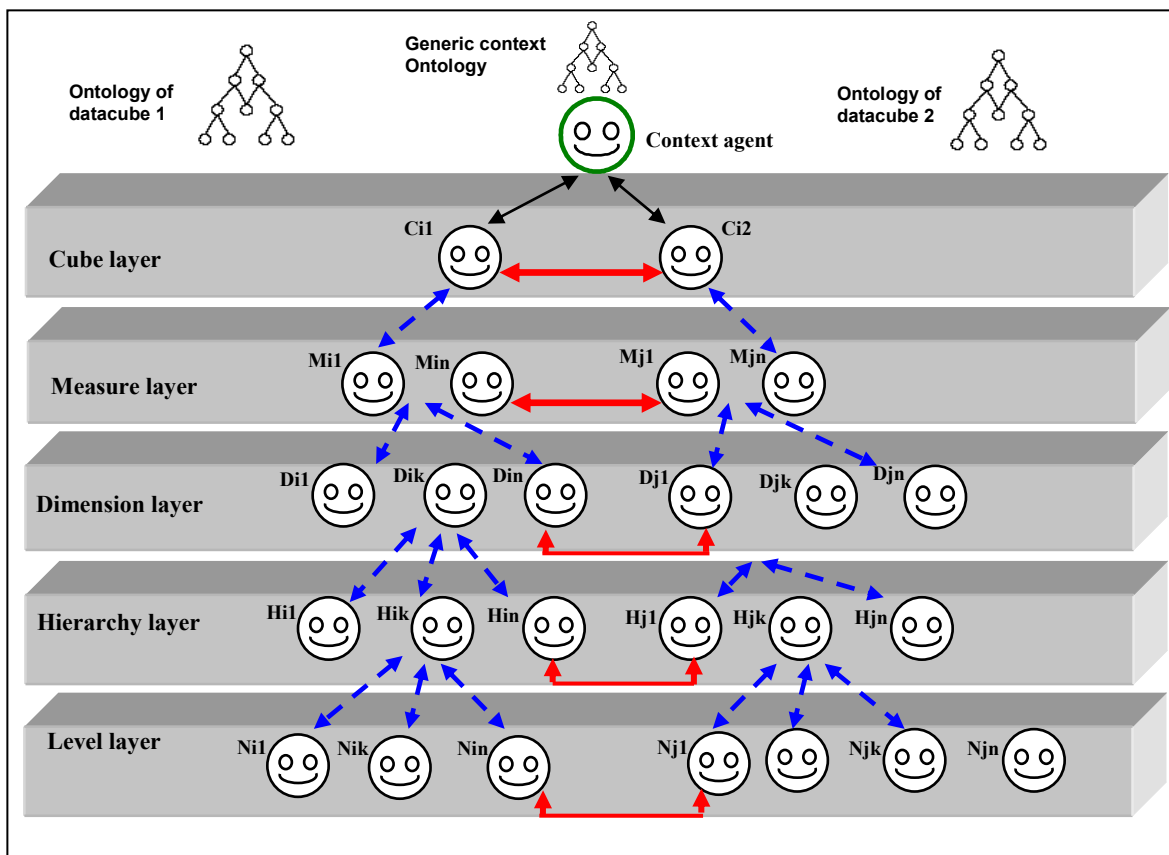


Figure 4.2: Conceptual framework between geospatial datacube agents.

The vertical communication allows an agent in a given layer to inform an agent in a lower layer of the same datacube about potential semantic heterogeneity (c.f., chapter 3). That is, an agent of a layer i of a datacube DI (A_{DIi}) informs an agent of a layer $i-1$ (A_{DIi-1}) of the same datacube about potential semantic problems that may occur during his/her/its

communication with another agent of a layer i of a datacube $D2$ (A_{D2i}). Such information would be important for the agent A_{D1i-1} to make appropriate decisions about the process of the interoperability (e.g., should he/she continue or withdraw from the process of interoperability?). For example, the *agent city* of a datacube $C1$ may decide not to communicate with the *agent ville* of a datacube $C2$ because the *agent region* of $C1$ (higher level agent) informed him/her about possible risky semantic problems that may cause bad consequences.

This hierarchical structure of the framework has the following advantages:

- It corresponds to the categorization of semantic problems proposed in chapter 3 (c.f., section 3.4). It allows agents to resolve the semantic heterogeneity occurred in each layer at once. This reduces the complexity of the overall semantic problems. That is, instead, for example, of assigning a single agent who is responsible for resolving semantic heterogeneity at the dimension layer and at its sub-layers (i.e., hierarchy and level layers), we assign three agents; each responsible for resolving semantic heterogeneity that may occur at each layer at once (i.e., dimension, hierarchy and level layers). As such, each agent deals with smaller, more manageable chunks of semantic heterogeneity.
- It allows a progressive resolution of semantic conflicts that may occur when interoperating geospatial datacubes. In fact, starting from the highest, and going down to the lowest level of each datacube, agents can resolve semantic conflicts that may occur at one level. At each level, agents can formulate comments and observations, and transmit them to the following level of the same geospatial datacube. For example, after communicating with a dimension agent of a datacube $D2$ ($A_{D2dimension}$), a dimension agent of a datacube $D1$ ($A_{D1dimension}$) can transmit a comment to the level agent of same datacube $D1$ ($A_{D1level}$) caution him/her to be careful when communicating with the level agent of $D2$ ($A_{D2level}$). This allows software agents to solve the problems of heterogeneity more efficiently with less time and effort (e.g., agents may suspend the resolution of problems before going too far within the hierarchical levels).

In our example of the Figure 3.8, we assign an agent to each datacube's level: for datacube C1, we assign an agent dimension that deals with datacube dimensions (*Time*, *Administrative region*, and *Age category*), an agent measure that deals with the measure *Distribution*, an agent hierarchy for the hierarchy $\langle \text{City}, \text{Province}, \text{Territory} \text{ and } \text{Country} \rangle$, and an agent level that deals with the levels of each datacube's dimension (e.g., the levels *City*, *Province*, *Territory*, and *Country* of the dimension *Administrative region*). For datacube C2, we assign an agent dimension that deals with datacube dimensions (*Period*, *Region* and *Forest stand*), an agent measure that deals with the measure *Fire zone*, an agent hierarchy for the hierarchies $\langle \text{City}, \text{Province}, \text{Territory} \text{ and } \text{Country} \rangle$, and $\langle \text{City}, \text{State} \text{ and } \text{Country} \rangle$, and an agent level that deals with the levels of each datacube's dimension (e.g., the levels *City*, *State*, *Province*, *Territory*, and *Country* of the dimension *Region*). Finally, we define a context agent that mediates the communication between these agents.

Agents representing the same level in different geospatial datacubes (e.g., dimension agent of C1 -to- dimension agent of C2) communicate horizontally. For example, dimension agent of datacube C1 (A_{dC1}) sends the name of the dimension name ("Administrative region") and the context associated with this dimension (e.g., "English language") to the dimension agent of datacube C2 (A_{dC2}). The latter searches for similar dimension names within his/her/its ontology. The search consists of comparing the semantics of the received data ("Administrative region") with the semantics of each dimension name of his/her/its knowledge base. A_{dC2} finds the dimension named "Region". Then, he/she asks the context agent to verify that the context associated with the received dimension name (e.g., "English language") is relevant with the context associated with the resulting dimension name. In this case, since both contexts are similar ("English language"), context agent sends the following message to A_{dC2} : "No context conflict has been noticed". Then, the A_{dC2} communicates vertically with the agent representing the level "Country" of the same datacube (i.e., datacube C2) by sending the following message: "No complex semantic conflict has been noticed. You can safely communicate".

In order to support data exchange between the agents of the proposed framework, we define a communication protocol that is based on a context agent and datacube agents (see Figure 4.3). In this protocol, each destination agent tries to resolve the semantic conflicts by

comparing the received data to his/her/its knowledge base (e.g., the ontology). Then, this agent sends a message to the context agent asking him/her to verify that the context associated with the received data is relevant with his/her/its context (i.e., receiver's context). The context agent verifies if there is a semantic conflict between the sender's context (i.e., production context) and the receiver's context (i.e., the context in which data is being used). If there is no conflict, the context agent gives the authorization to the current data exchange. Else, he/she sends a warning to the destination agent to make him/her aware of context conflicts. Then, destination agent sends a message to the agent representing a lower level of the same datacube making him/her aware of potential semantic conflicts that may affect his/her/its communication. In this case, the agent of the lower level should make a decision about whether to continue or suspend the communication process. If he/she decides to continue, he/she should keep in mind that there may be undesirable consequences (e.g., data may not be re-usable in his/her/its context).

We should note that the context agent stores any information about the context in what we call *context knowledge*. Context knowledge will be used to resolve the semantic conflicts that may occur during future communication between geospatial datacubes agents.

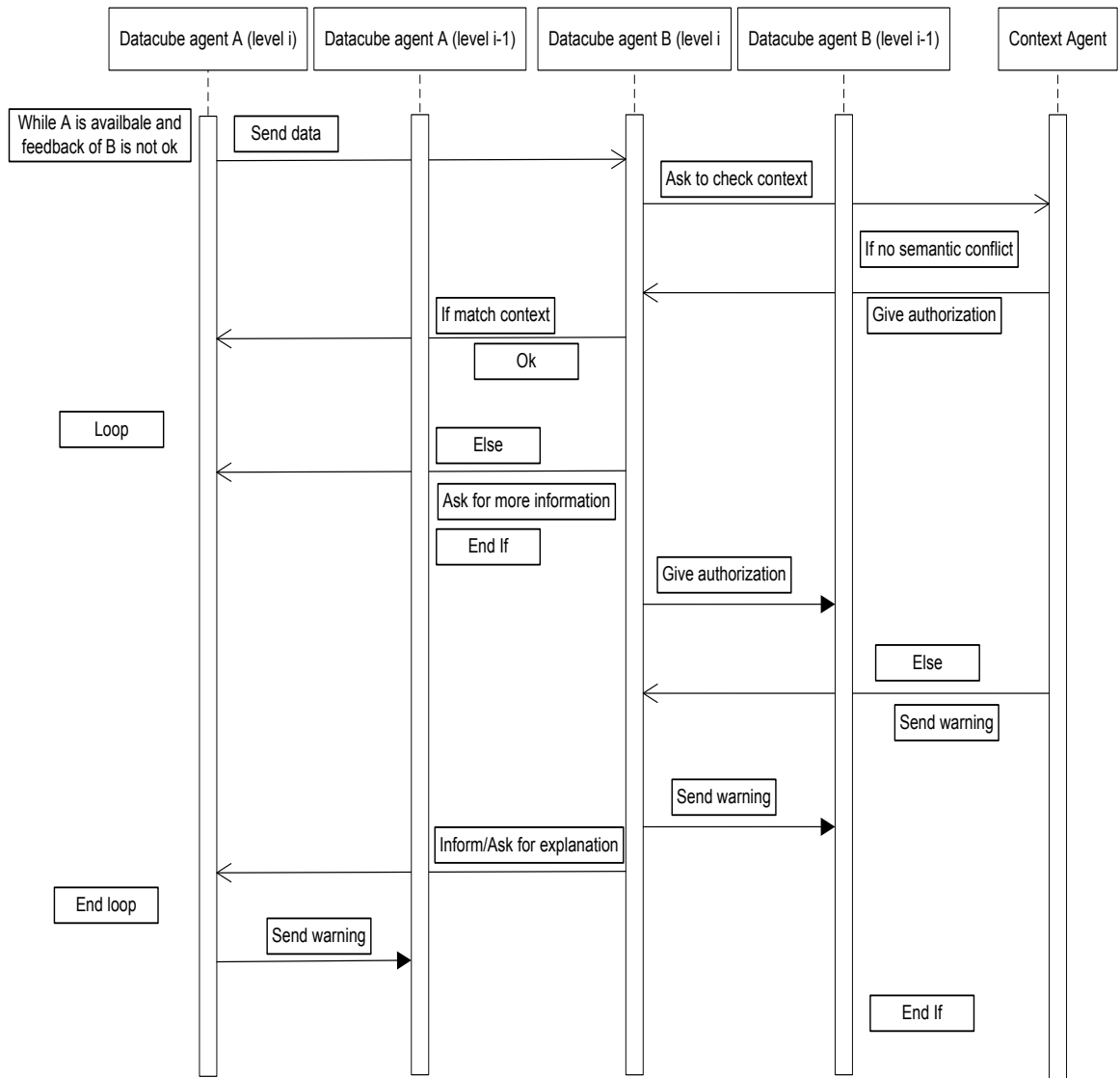


Figure 4.3: The communication protocol for the proposed conceptual framework

In order to communicate, agents usually need to compare the elements of their respective ontologies. Comparing ontologies allows determining the semantic similarity between the content of heterogeneous geospatial datacubes.

4.3 Resolving semantic heterogeneity in geospatial datacubes

4.3.1 Using the GsP notion

In order to deal with semantic problems at each layer of the proposed framework, each agent uses two ontologies: one contains the concepts of geospatial datacubes and another represents the context related to the datacube's content. Ontologies have been considered as the backbone technology used to solve semantic heterogeneity in the interoperability of information systems, and to interpret exchanged data (Gruber 1993, Guarino 1998, Kuhn 2005). The process of resolving semantic heterogeneity involves a chain of comparing different concept representations (e.g., the destination agent compares the received concept representation with a set of concept representations stored in its own ontology). The process usually ends when the resulting data are familiar to the interpreting agent. This might involve successive generalizations (using one or more generic ontologies) until a common concept is found (Bailin and Truszkowski 2002).

As seen in chapter 2, many approaches have been using ontologies to resolve the semantic heterogeneity that may occur between heterogeneous information systems. Examples of research works are: the Semantic Formal Data Structure model (Bishr 1997), the Matching Distance model (Rodriguez 2000), the Isis approach (Benslimane 2001), geosemantic proximity (Brodeur 2004), the G-Match tool (Hess 2006) and the similarity-based information retrieval approach (Janowicz et al. 2008). After reviewing these works, we chose the geosemantic proximity approach (GsP) proposed by Brodeur (2004) that evaluates the semantic similarity between geospatial concepts (i.e., similarities between their intrinsic and extrinsic properties). We use GsP to compare different representations of data exchanged between agents at each layer of our conceptual framework. This includes using GsP to resolve the context heterogeneity related to exchanged data (e.g., the heterogeneity of geospatial referencing systems as defined in the category *level-to-level heterogeneity*, cf. section 3.4).

The choice of the geosemantic proximity approach is motivated by different reasons:

1. Geosemantic proximity is based on human-like communication which we believe is appropriate for the interoperability process.

2. Geosemantic proximity deals with geometric aspects of data. It measures the semantic similarity of geospatial data.
3. Geosemantic proximity tool was developed and tested in our research group. Consequently, it was possible to acquire the source code, and adapt and extend it to support the interoperability of geospatial data cubes.

The GsP evaluates qualitatively the semantic similarity between geospatial concepts. It compares the inherent properties of one concept with another. These properties are classified in two types: intrinsic and extrinsic. Intrinsic properties provide the literal meaning of a concept. They consist of the identification, the attributes, the attribute values, the geometries, the temporalities, and the domain of the concept. Extrinsic properties are properties that are subject to external factors (e.g., behaviours and relationships). The semantic of a geospatial concept is defined by the union of intrinsic and extrinsic properties. Then, the GsP of two concepts can be defined by the intersection of their respective properties. It results in a four-intersection matrix when consolidated with intrinsic and extrinsic properties (Brodeur 2004). Each component of the matrix can be evaluated empty (denoted by *f* or false) or not empty (denoted by *t* or true). Accordingly, 16 predicates were derived which are consistent with Allen's and Egenhofer's well known approaches of temporal and spatial topological relationships (Allen 1981, Allen 1983, Egenhofer 1993, Egenhofer and Franzosa 1991).

The predicates are: GsP_ffff (or disjoint), GsP_ffft, GsP_fftt (or contains), GsP_tfft (or equal), GsP_ftft (or inside), GsP_tfft (or covers), GsP_ttft (or coveredBy), GsP_fttt (or overlap), GsP_tttt, GsP_tfff (or meet), GsP_tftf, GsP_tttf, GsP_tfff, GsP_fttf, GsP_fttf, GsP_ftff (Brodeur 2004).

In the following, we propose an extension to the GSP notion in order to support the semantic interoperability between geospatial data cubes. The extension, called MGsP, aims to give the possibility to dig into and resolve semantic heterogeneity related to key notions of the multidimensional paradigm.

4.3.2 MGsP: Extending the GsP notion to stress the semantics of geospatial datacubes elements

The hierarchical structure of dimensions and the dependencies between dimensions and measures induce several semantic conflicts specific to the multidimensional data. Notably, the semantic heterogeneity of aggregation of dimension levels, semantic heterogeneity of measure function, and the semantic heterogeneity of hyper-cell (Salehi 2009, cf. chapter 2) present a particular obstacle when interoperating different geospatial datacubes. Thus, we intend to provide agents (datacube agents and the context agent) a way to check for and resolve semantic heterogeneities specific to those particular elements.

For that, we propose an extension of the GsP notion to include the comparisons of basic multidimensional concepts such as the semantic of aggregation and the semantic of hyper-cellability. The goal of this extension (called multidimensional geosemantic proximity: MGsP) is to give agents the possibility to focus on the heterogeneity of multidimensional data by digging into more details about the semantic aspects of important notions of the multidimensional paradigm (e.g., aggregation, measure function, and hyper-cellability). As such, agents can focus on the multidimensional characteristics and make appropriate decisions with regards to their semantic similarity. Accordingly, we define three attributes to specialize the GsP: *dimension aggregation*, *measure function* and *hyper-cellability*. We should note that we chose these attributes, which related to dimensions and measures, as examples to illustrate the usefulness of the GsP extension for the interoperability between geospatial datacubes. This choice is motivated by the wide use of these attributes in the multidimensional paradigm. One can add other attributes if needed.

As in GsP, our methodology for qualitatively evaluating the semantic similarity consists of identifying the relations between decisional elements of heterogeneous geospatial datacubes according to their hyper-cellability, measure function and aggregation dimension. The semantics of the multidimensional elements (dimension or measure) is evaluated as the union of the properties related to the measure function (or dimension aggregation) and the properties related to the hyper-cellability.

Let:

M: a measure

D: a dimension

MInP: set of intrinsic multidimensional properties (for measure: $MInP = MInP_M$, whereas for dimension: $MInP = MInP_D$).

Where: $MInP_M$ is the set of properties related to the measure function. The function is considered as intrinsic property since it refers to the meaning of the measure, and

$MInP_D$ is the set of properties related to the aggregation. The aggregation is considered as intrinsic property since it refers to the meaning of the dimension.

MExP: set of properties related to the hyper-cellability. The hyper-cellability refers to the dependencies of measures with dimensions. Thus, it is considered as extrinsic property for both dimensions and measures.

MS_M : Multidimensional semantics of measure.

MS_D : Multidimensional semantics of dimension.

Then: $MS_M = MInP_M \cup MExP$

$MS_D = MInP_D \cup MExP$

The multidimensional geosemantic proximity (MGsP) is determined according to the intersection between the semantic of two elements (E1 and E2) of heterogeneous datacubes.

Let:

MS_{E1} : Multidimensional semantics of E1,

MS_{E2} : Multidimensional semantics of E2,

MGsP (E1, E2): Multidimensional Geosemantic proximity between E1 and E2.

Then: $MGsP(E1, E2) = MS_{E1} \cap MS_{E2}$

Accordingly, we define a 4-Intersection matrix containing the following four topological sub-relations (see Figure 4.4). In this matrix:

$MInP_E \in InP_E$ (the properties related to the measure function (or to the aggregation) belong to the intrinsic properties defined in GsP).

$MExP_E \in ExP_E$ (the properties related to the hyper-cellability belong to the extrinsic properties defined in GsP)

Thus, MGsP's matrix is a specialization of the one defined in the GsP, allowing agents to dig into more details of the multidimensional aspects of geospatial datacubes.

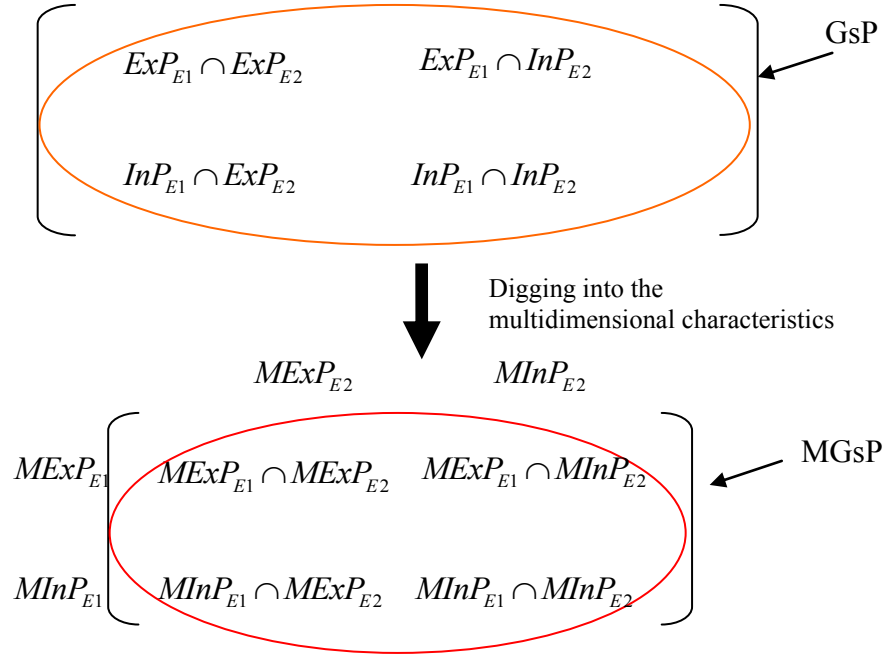


Figure 4.4: 4-intersection multidimensional matrix as a specialization of the GsP.

Since we consider the *function of measure* as an attribute of the measure's intrinsic properties, whereas the *hyper-cellability* as an attribute of the measure's extrinsic properties, we represent a 4-Intersection matrix for measure as follows:

$$\begin{array}{cc}
& \text{hyper_cell}_{M2} & \text{measure_function}_{M2} \\
\begin{array}{c} \text{hyper_cell}_{M1} \\ \text{measure_function}_{M1} \end{array} & \left(\begin{array}{cc} \text{hyper_cell}_{M1} \cap & \text{hyper_cell}_{M1} \cap \\ \text{hyper_cell}_{M2} & \text{measure_function}_{M2} \\ \text{measure_function}_{M1} \cap & \text{measure_function}_{M1} \cap \\ \text{hyper_cell}_{M2} & \text{measure_function}_{M2} \end{array} \right)
\end{array}$$

Since we consider the *aggregation* as an attribute of the dimension's intrinsic properties, whereas the *hyper-cellability* as an attribute of the dimension's extrinsic properties, we define the following 4-Intersection matrix for dimension:

$$\begin{array}{cc}
& \text{hyper_cell}_{D2} & \text{aggregation}_{D2} \\
\begin{array}{c} \text{hyper_cell}_{D1} \\ \text{aggregation}_{D1} \end{array} & \left(\begin{array}{cc} \text{hyper_cell}_{D1} \cap & \text{hyper_cell}_{D1} \cap \\ \text{hyper_cell}_{D2} & \text{aggregation}_{D2} \\ \text{aggregation}_{D1} \cap & \text{aggregation}_{D1} \cap \\ \text{hyper_cell}_{D2} & \text{aggregation}_{D2} \end{array} \right)
\end{array}$$

As in GsP, the comparison of properties between two elements (measures or dimensions) of heterogeneous datacube could be evaluated empty (denoted by \emptyset or f) and non-empty (denoted by $\neg\emptyset$ or t) expressing respectively that none or some properties are common. This leads to 16 (i.e., 2^4) possible MGsP predicates for each matrix (see Figure 4.5).

If $\text{hyper_cell}_{M1} \cap \text{measure_function}_{M2}$ is $\neg\emptyset$ (\emptyset), it indicates that the measure function of M2 fits (respectively does not fit) to the hyper-cellability of M1.

If $\text{hyper_cell}_{D1} \cap \text{aggregation}_{D2}$ is $\neg\emptyset$ (\emptyset), it indicates that the aggregation of D2 fits (does not fit respectively) to the hyper-cellability of D1.

In Figure 4.5, MGsP predicates are organized in four distinct categories according to four characteristics: common MExP and common MInP, common MExP and no common MInP, no common MExP and no common MInP, and no common MExP and common MInP.

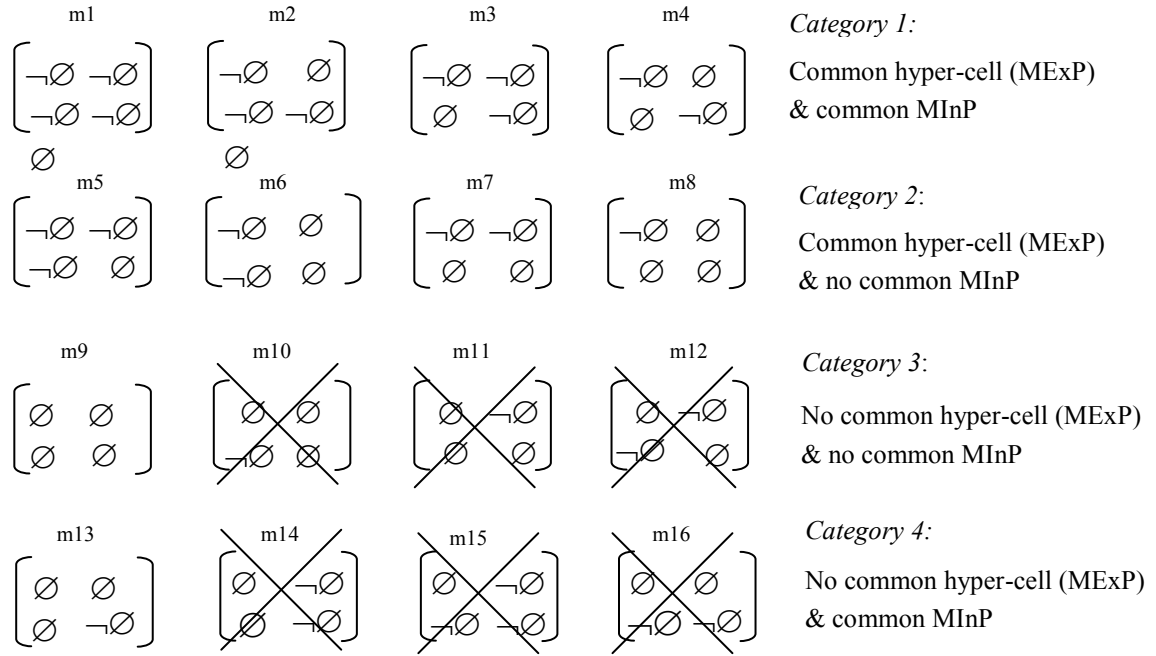


Figure 4.5: 16 possible MGsP predicates.

1) *Category 1:*

This category refers to the case where both the functions and the hyper-cellability of the heterogeneous measures are common. This category includes four possible matrixes:

a. MGsP_{tttt} (E1, E2)

$$\begin{bmatrix} \neg\emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset \end{bmatrix}$$

In this particular case the function of M1 (respectively M2) fits the hyper-cellability of M2 (respectively M1).

b. MGsP_{tftt} (E1, E2)

$$\begin{bmatrix} \neg\emptyset & \emptyset \\ \neg\emptyset & \neg\emptyset \end{bmatrix}$$

In this case the function of M1 fits the hyper-cellability of M2. However, the function of M2 does not fit the hyper-cellability of M1.

c. $\text{MGsP_tfft}(E1, E2)$

$$\begin{bmatrix} \neg\emptyset & \neg\emptyset \\ \emptyset & \neg\emptyset \end{bmatrix}$$

In this particular case the function of M1 does not fit the hyper-cellability of M2. However, function of M2 fits the hyper-cellability of M1.

d. $\text{MGsP_tfft}(E1, E2)$

$$\begin{bmatrix} \neg\emptyset & \emptyset \\ \emptyset & \neg\emptyset \end{bmatrix}$$

In this case the function of M1 (respectively M2) does not fit the hyper-cellability of M2 (respectively M1).

2) Category 2:

This predicate refers to the case where the hyper-cellabilities of the heterogeneous measures are common, whereas the functions are dissimilar. This category includes four possible matrixes:

a. $\text{MGsP_tttf}(E1, E2)$

$$\begin{bmatrix} \neg\emptyset & \neg\emptyset \\ \neg\emptyset & \emptyset \end{bmatrix}$$

In this particular case the function of M1 (respectively M2) fits the hyper-cellability of M2 (respectively M1).

b. $\text{MGsP_tftf}(E1, E2)$

$$\begin{bmatrix} \neg\emptyset & \emptyset \\ \neg\emptyset & \emptyset \end{bmatrix}$$

In this case the function of M1 fits the hyper-cellability of M2. However, function of M2 does not fit the hyper-cellability of M1.

c. $\text{MGsP_ttff}(E1, E2)$

$$\begin{bmatrix} \neg\emptyset & \neg\emptyset \end{bmatrix}$$

$$\emptyset \quad \emptyset$$

In this particular case the function of M1 does not fit the hyper-cellability of M2. However, function of M2 fits the hyper-cellability of M1.

d. $\text{MGsP_tfff}(E1, E2)$

$$\begin{bmatrix} \neg\emptyset & \emptyset \\ \emptyset & \emptyset \end{bmatrix}$$

In this case the function of M1 (respectively M2) does not fit the hyper-cellability of M2 (respectively M1).

3) *Category 3:*

$\text{MGsP_ffff}(E1, E2)$:

$$\begin{bmatrix} \emptyset & \emptyset \\ \emptyset & \emptyset \end{bmatrix}$$

This predicate refers to the case where both the functions and the hyper-cellability of the heterogeneous measures are dissimilar. We should note that in this case the function of M1 (respectively M2) should not fit the hyper-cellability of M2 (respectively M1). Accordingly, we do not consider the matrix m10, m11 and m12 (see Figure 4.5).

4) *Category 4:*

$\text{MGsP_ffft}(E1, E2)$:

$$\begin{bmatrix} \emptyset & \emptyset \\ \emptyset & \neg\emptyset \end{bmatrix}$$

This predicate refers to the case where the hyper-cellabilities of the heterogeneous measures are dissimilar, whereas the functions are similar. We should note that, in this case, since there is no intersection between the hyper-cellabilities, the function of M1 (respectively M2) should not fit the hyper-cellability of M2 (respectively M1). Accordingly, we do not consider the matrix m14, m15 and m16 (see Figure 4.5).

Ten resulting predicates are then defined for the MGsP of measures, which are: MGsP_tttt, MGsP_tfft, MGsP_ttft, MGsP_tfft, MGsP_tttf, MGsP_tftf, MGsP_tfff, MGsP_ftff, and MGsP_ffft.

Similarly, we define the predicates for the multidimensional geosemantic for the dimensions. The resulting predicates are the same ones as those for the measure element.

Using such attributes (hyper-cell, dimension aggregation, and measure function) agents can have a better idea about the semantic similarity of multidimensional concepts, and hence can make appropriate decisions about resolving the semantic heterogeneity that may occur between the elements of different geospatial datacubes. For example, if the functions of two semantically heterogeneous measures (e.g., *density* in a datacube C1 and *concentration* in datacube C2) are completely different, agents may consider these measures are dissimilar even if they have other common characteristics (e.g., used for the same subject of analysis, represented with the same precision and having the same scale). As such, based on ontologies, the MGsP can facilitate the interpretation of the content of multidimensional geospatial datacubes involved in the interoperability process.

However, as for the MGsP notion, ontology-based interpretation still remains very difficult. This is basically due to the fact that it is sometimes impossible to capture and reason about geospatial data semantics based solely on ontologies.

4.4 Limits of the ontology-based interpretation of geospatial concepts

Despite interesting research works, including the geosemantic proximity approach, ontology-based interpretation, which basically consists of matching heterogeneous concept representations, still suffers from problems related to the quality of matching results. That is, the matching results are only an unreliable estimation of the similarity degree (Bouquet et al. 2005, Eckert et al. 2009). Such estimation affects the accuracy of resolving semantic heterogeneity and hence raises questions about the effectiveness of the interpretation of the content of heterogeneous geospatial datacubes.

Moreover, in the semantic interoperability between geospatial datacubes, data may be given an interpretation which is different from the intended meaning, or wrongly adapted to the current use (i.e., the use for which the interoperability is carried out). This results from the fact that data description may refer to unintended interpretations (i.e., extra interpretations that were not intended by the data producer). Accordingly, in different contexts of interoperability (e.g., different purposes or different skills of stakeholders), data may have different interpretations. These interpretations may have different degrees of closeness to the intended meaning, and different degrees of relevance to a given context (i.e., a context for which the interoperability is carried out). Figure 4.6 shows that the interpretation in a context *A* is closer to the intended meaning than the one in context *B*.

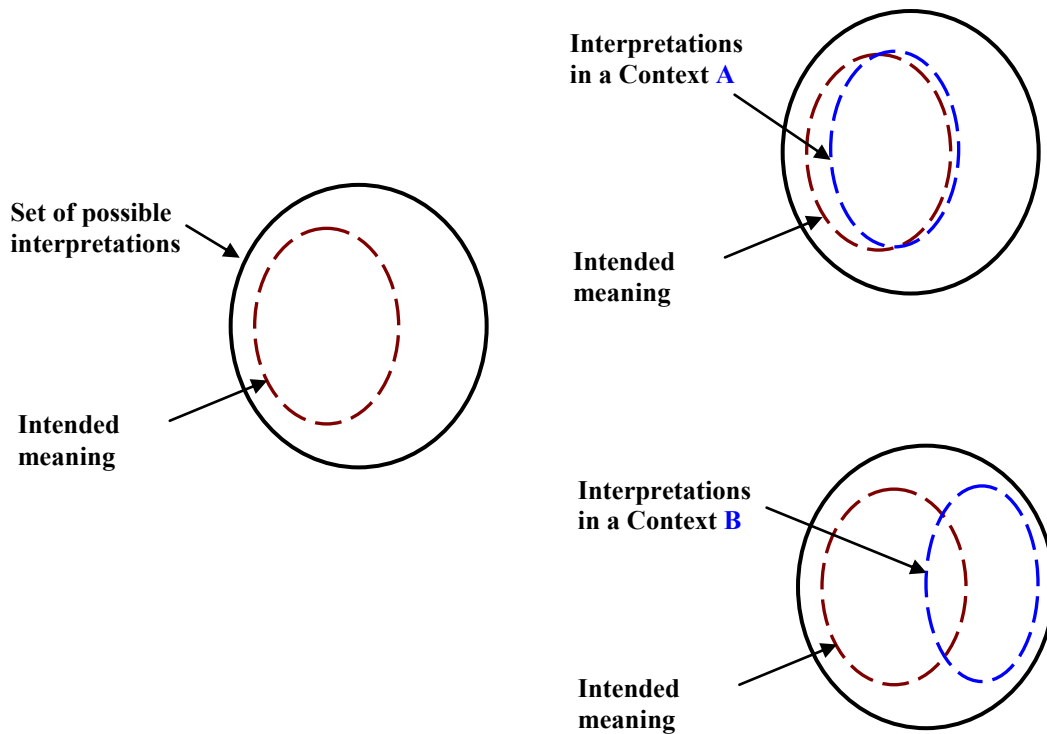


Figure 4.6: Different interpretations in different contexts.

Usually, agents are uncertain about either how much their interpretation is close to the intended meaning or how much this interpretation is complete and relevant with regards to the current context of use, or about both. The uncertainty about such degree of closeness and/or of relevance, we identify it as *semantic uncertainty*, during the interoperability process may affect the decisions of stakeholders (human interveners or software agents).

In semantic interoperability between geospatial datacubes, semantic uncertainty may occur at each one of the five levels of heterogeneity previously defined: cube-to-cube, measure-to-measure, dimension-to-dimension, hierarchy-to-hierarchy, and level-to-level. At each level, semantic uncertainty may have impact upon the decisions of stakeholders. That is, semantic uncertainty is considered as another barrier, besides semantic heterogeneity, for the interoperability of geospatial datacubes, and thus affects the strategic decision-making for which such interoperability is carried out.

The impact of such uncertainty is more apparent in cases where stakeholders need to make decisions about whether considering one or more elements among others. In such cases, and especially in critical situations (e.g., forest fire in adjacent countries), we cannot afford to be uncertain about data interpretation. Such cases include:

- a. the choice between different elements of the heterogeneous geospatial datacubes to be interoperated (e.g., to choose between two dimensions *Administrative region* and *Region* to create a new one),
- b. whether to add a datacube element of one datacube to another (e.g., whether to insert the dimension *Age category* of the datacube *Manpower* in the datacube *Human resources*), or
- c. whether to consider several heterogeneous elements (e.g., whether to consider and merge the levels *Province* and *State* to create a new one, or to consider the dimension *Province* of the datacube *Canada 1980* to update the dimension *Territory* of the datacube *Canada 2000*).

Consequently, evaluating the degree of semantic uncertainty related to the interpretation of geospatial datacubes elements is useful to avoid considering them when their use is associated with unacceptable risks of data misinterpretation. Therefore, evaluating such uncertainty should determine whether the associated risks are either acceptable or unacceptable because they present significant harm to the interoperability process. In chapter 5, we discuss the risks of data misinterpretation related to semantic interoperability between geospatial datacubes.

We should notice that we cannot deal with all aspects of semantic uncertainty by relying only on machines. In fact, it is highly difficult for software agents (machines) to automatically capture and interpret all possible concepts and relationships that may exist in a particular application (e.g., suggestive concepts and transitive dependencies). Moreover, as discussed, ontologies, which have been considered as the backbone of the semantic interoperability, do not provide a perfect solution for resolving semantic heterogeneity of geospatial datacubes (nor for other information systems), and for dealing with all the causes of semantic uncertainty. Consequently, it appears that human intervention is essential to deal with the risks related to semantic uncertainty in the context of interoperability between geospatial datacubes.

4.5 Human intervention to support semantic interoperability between geospatial datacubes

It may be extremely difficult (even impossible) for machines to deal with all semantic problems to appropriately interpret data in the interoperability process. For example, consider the following message from Agent A to Agent B: “restaurants near Chateau Frontenac”. In order to appropriately interpret the vague concept *near*, we may need to be aware of Agent A’s attitudes and preferences. Agent A may consider, for example, that restaurants within 5 kilometres of Chateau Frontenac to be “near”. Context information (e.g., agent’s attitudes, preferences, available methods of traveling, time of the day, traffic info) may be relevant to appropriately interpret such data. Taking into account such information requires an extensive knowledge and judgement which are unique to humans and cannot be completely automated (Swanson and Smalheiser 1997).

To illustrate this point of view, we revisit the example of the datacube (see Figure 3.8). In this datacube, the generalization used to insert data in the level *Province* from the level *City* of the dimension *Administrative region* was not fully automated and required the intervention of a human expert to carry out different tasks (e.g., elimination or replacement of geospatial objects, displacement of others, simplification of shapes, and change of topological relationships). In his/her intervention, the expert incorporates his/her subjective judgment and preferences. Receiving generalized data of the level *Province* from an agent

representing datacube C1, an agent representing datacube C2 (machine) cannot be aware of all information surrounding the generalization process (especially information related expert's judgment and preferences) (Abbott 2004), and hence cannot appropriately interpret the received data. Thus, there is a need for human intervention to take into account such complex information (e.g., suggestive and vague concepts), and appropriately interpret data. Human stakeholders can play the role of context agent proposed in our framework when it is required (see Figures 4.1 and 4.2).

In order to guide human stakeholders to appropriately interpret the content of geospatial datacubes, we propose a semantic model, called *SemEL*. This model enables to explicitly represent data meaning, and help human stakeholders to intuitively interpret data according to ontology, spatio-temporal characteristics, and context. Accordingly, *SemEL* contains an ontology sub-model, a spatio-temporal sub-model, and a context sub-model. The ontology sub-model has three dimensions (i.e., definitions, assumptions, and properties) and a fact table that contains ontology description (*Ontology_Desc*). The spatio-temporal sub-model has four dimensions (i.e., positions, geometries, graphics and times) and a fact table that has the spatio-temporal description of datacubes concepts (*Spatio-temporal_Desc*) as its unique measure. The context sub-model has four context dimensions (*Goal Context*, *Domain Context*, *Dataset Context*, and *Concept Context*) and a fact table that contains context description (*Context_Desc*). This contribution was the main subject of the paper titled "A Conceptual Framework to Support Semantic Interoperability of Geospatial Datacubes" (Sboui et al. 2007). An extraction of this article can be found in Annex B of this thesis.

In the ideal case, human stakeholder would be a group of datacube designers. Moreover, at least one of the members of the group would have experience in dealing with geospatial data, in business intelligence (strategic decision process) and in other fields that may surround the interpretation and the use of geospatial datacubes contents (e.g., ethical and legal aspects). Moreover, experts may rely on their experience and knowledge to support end-users in describing their needs, in defining the circumstances that surround their applications, and in adapting data to their specific use.

In order to guide both humans and software agents to deal with the risks of data misinterpretation, we propose, in the next chapter, a risk management approach that aims to identify, evaluate and make appropriate decisions.

4.6 Conclusions

Interoperability has been viewed as an efficient paradigm to reuse geospatial data. The interoperability between geospatial datacubes has specific characteristics and particular semantic conflicts. Although one could do it using traditional transactional solutions, the efficiency can be improved with an enriched framework that explicitly supports geospatial datacubes. In this chapter, we proposed a conceptual framework to support semantic interoperability between these datacubes. The framework is based on software agents (datacube agents and a context agent). Datacube agents represent geospatial datacubes to be involved in the interoperability process, and are engaged in a process of human-like communication. They are defined according to five communication layers that correspond to the five conceptual levels of geospatial datacube; from the more general level to the most detailed level (cube, measure, dimension, hierarchy, and level), and to the five categories of semantic heterogeneity of geospatial datacubes (cube-to-cube, measure-to-measure, dimension-to-dimension, hierarchy-to-hierarchy, and level-to-level). In order to facilitate data exchange between datacube agents, context agent mediates between them to make them aware of potential mismatch or inappropriateness of context information.

The proposed conceptual framework defines two types of communication: 1) horizontal communication between agents representing the same level belonging to different geospatial datacubes, and 2) a vertical communication between agents of the same datacube at different levels (e.g., dimension of datacube A -to- level of datacube A). The hierarchical structure of the proposed framework corresponds to the categorization of semantic problems proposed in chapter 3 (c.f., section 3.4). As such, it allows agents to resolve the semantic heterogeneity occurred in each category at once; starting from the highest and going down to the lowest level of each datacube. This reduces the complexity of the overall semantic problems.

In order to resolve the semantic heterogeneity that may occur between heterogeneous geospatial data cubes, we opted for the geosemantic proximity approach (GsP). Moreover, and in order to deal with particular semantic heterogeneities of geospatial data cubes, we proposed an extension of the GsP approach (MGsP) to stress the semantic of basic multidimensional concepts such as the semantic of aggregation, the semantic of measure function and the semantic of hyper-cellability. The objective of this extension is to give agents the possibility to dig into more details about the semantic heterogeneity of important notions of the multidimensional paradigm.

Despite various researches to measure the semantic similarity between geospatial concepts (including the MGsP notion), automatic ontology-based matching still suffers from problems related to the quality of matching results. Moreover, data description may have different interpretations which are more or less relevant to the intended meaning (semantic uncertainty). These result in risks of data misinterpretation. Existing approaches of interoperability tried to deal with such risks in a non-systematic manner (i.e., not based on predefined and ordered criteria).

Machines cannot deal with the risks of data misinterpretation by itself. In fact, they cannot capture and interpret information that requires an extensive knowledge and judgement which are unique to humans. Accordingly, human intervention still remains needed to capture and interpret some concepts and relationships that may exist in the interoperability between geospatial data cubes. However, humans normally do not have the ability to deal with a large amount of exchanged data. Hence, both human stakeholders and software agents still need to be supported to identify the risks of data misinterpretation, and making decisions about such risks.

In the next chapter, we propose a risk management approach to deal with the risks of data misinterpretation in the interoperability between geospatial data cubes. This approach intends to support the intervention of both machines and humans to control these risks.

Chapter 5: Fine-tuning the semantic interoperability between geospatial datacubes with a risk management approach based on the fitness-for-use of conceptual models

5.1 Introduction

In the previous chapter, we saw that, with the emergence of software agents, semantic interoperability has been viewed as the technical analogue to human communication (Sboui et al. 2007, Kuhn 2005, Brodeur 2004). According to this view, we proposed a conceptual framework for the interoperability between geospatial datacubes. The framework defines two types of agents: datacube agents and a context agent. In order enable these agents to resolve the semantic heterogeneity, we use and extend the GsP approach in developing MGsP. However, the interoperability between geospatial datacubes still remains vulnerable to the risks of data misinterpretation. These risks cannot be ignored and should be deal with carefully. In fact, geospatial datacubes are developed for strategic decision purposes. Strategic decisions made on the basis of inappropriate interpretations of data may lead analysts to have inappropriate judgment, and thus to make faulty decisions. In order to deal with the risks of data misinterpretation, we need to identify them, assessing their degree of severity and making decisions about them.

In this chapter, we discuss such risks and we examine their potential causes and consequences. Then, we propose a risk management approach that allows identifying and assessing the risks of data misinterpretation in a systematic manner based on the quality of geospatial datacubes models (production context and schemas). The approach consists of evaluating the quality of the conceptual models of datacubes involved in the interoperability process, and of a framework that aims to facilitate decision-making in responding to such risks.

In the next section, we discuss the risks of data misinterpretation related to semantic interoperability in general and to the one between geospatial datacubes in particular. In

section 5.3, inspired from the risk management in the field of project management, we propose a set of indicators to identify and evaluate the risks of data misinterpretation during the interoperability process between geospatial datacubes. In Section 5.4, we propose a framework to support a stakeholder (human or system agent) to respond to these risks. In section 5.5, we provide an example of application to illustrate the proposed approach. We conclude this chapter in Section 5.6.

5.2 Semantic interoperability between geospatial datacubes and risks of data misinterpretation

The imperfect resolution of semantic heterogeneity and the semantic uncertainty related to data interpretation result in **risks of data misinterpretation**. That is, **the probability of faulty interpreting data or being uncertain about the relevance of data interpretation with regard to its use, and the consequences that may arise as a result**. Existing approaches of interoperability tried to deal with such a risk in a non-systematic manner (i.e., not based on predefined and ordered criteria). As a result, stakeholders (human interveners or software agents) have to put extensive time and effort into identifying the risks related to data misinterpretation.

In our work, we intend to deal with such risks at each layer of the proposed conceptual framework of interoperability based on predefined and ordered criteria. In sections 5.3 and 5.4, we propose a method to deal with the risks of data misinterpretation based on well established paradigm for managing the risks within project management. But first, in this section, we review the notion of the risk and risk management, and we discuss the risks of data misinterpretation related to semantic interoperability in general and to the one between geospatial datacubes in particular.

5.2.1 Risk and risk management

There are various understandings of the term “risk” depending on the circumstances in which it is used. However, this term usually refers to the possibility that an undesirable outcome may occur as a result of an event. IEC considers the notion of risk as a

“combination of the probability of occurrence of harm and the severity of that harm” (IEC 2000).

But how can we qualify the result of an event as undesirable (or harm)? Technical approaches consider “harm” as physical dysfunction or error. Psychological approaches consider subjective judgment and preferences to determine if the outcome of an event as “harm” (Renn 1998). For example, downhill skiing may be viewed by some people as undesirable while perceived as desirable by others (Machlis and Rosa 1990).

Risk management refers to the process of reducing the risk to a level considered acceptable by an individual or an organization (Morgan 1990, Renn 1998). Managing risks consists basically of four phases: 1) identifying the risks, 2) assessing the risks (i.e., determining their probability of occurrence and their degree of harm), 3) taking proper actions to reduce the risk, and 4) documenting the previous phases (Machlis and Rosa 1990, ISO 2009).

But what risk level is considered acceptable? (i.e., How safe is safe enough?). The answer to this question depends on the circumstances that surround each case. For example, the acceptable risk of geospatial data quality deficiencies depends on the nature of application and people’s attitude and readiness to deal with risks (e.g., constructing a bridge versus looking for touristic attractions). Thus, there is no single “How safe is safe enough?” problem (Derby 1981).

5.2.2 Overview of risks of data misinterpretation in geospatial datacubes interoperability

Interpreting data as they were originally intended, and adapting it to the current context is the main aim of a destination agent. However, this aim is very difficult or even impossible to achieve since data have no built-in intrinsic signification (Schramm 1971, Bédard 1986). Consequently, the receiver may interpret data inappropriately.

For example, Agent *A* (representing a geospatial dataset developed in the province of Ontario using the English language) communicates a message about the member of the concept *River* (“The river 2 intersects with the river 1”) to Agent *B* (representing a geospatial dataset developed in the province of Quebec using the French language) (see

Figure 5.1). The concept *River* may be interpreted by Agent *B* (a French speaker) as 1) a stream of water that flows into a sea or as 2) a stream of water that flows into other water bodies. Receiving the data from Agent *A* (1), Agent *B* may wonder about the meaning of the concept *River* (i.e., is it a stream of water that flows into a sea (i.e., *Fleuve* in French) or a stream of water that flows into other water bodies (i.e., *Rivière* in French)?). That is, there is a risk that Agent *B* misinterprets the received data.

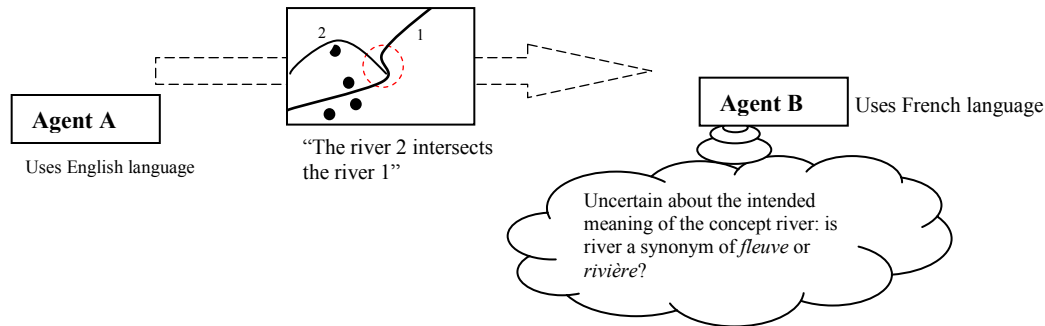


Figure 5.1: Risk of misinterpreting data (“The river 2 intersects with the river 1”).

Data misinterpretation may lead to a lack of understanding of real-world phenomena which may harm the reuse of geospatial data. The risks of data misinterpretation are even more significant in the context of the interoperability between geospatial datacubes. In fact, misinterpreting the content of geospatial datacubes may cause inappropriate judgments, and unwarranted inferences about some aspects of the problem to be solved, resulting in bad strategic decisions.

Dealing with the risk of data misinterpretation in geospatial datacubes is more complicated than dealing with it in transactional databases, due to the fact that:

- in geospatial datacubes, the initial transactional data have undergone complex ETL (Extract, Transform and Load) procedures that lead to the creation of analytical data and can impact the meaning of its resulting content. This adds to the fact that data are generally collected from other heterogeneous sources having themselves undergone complex procedures (see (Bédard 1986) for a detailed explanation). During the transformation phase, some interpretations may be formed and several rules, functions and decisions may be applied to the

collected data in order to fit business needs (e.g., modifying the terms used to be understood by users in business analysis).

- in geospatial datacubes, data are also aggregated or summarized using different methods. This aggregation adds another level of complexity of interpretation. In fact, in order to interpret data, users of geospatial datacube may need to understand first the method used for aggregating data since the same source data may be summarized using different methods. For example, the buildings surrounded by four streets can be aggregated to form what we call a building block if the density is higher than a given threshold. In order to truly understand the meaning of such a polygon, users need to know first the criteria used for the aggregation. Also, as shown in Figure 2.11, due to type of generalization being carried out, the same real-world phenomena may be represented differently (Figure 2.11 (a2) is interpreted as a block with 11 buildings, while Figure 2.11 (b2) is interpreted as a block with 8 buildings).
- geospatial datacubes promote a frequent human intervention especially when extracting, cleansing, aggregating, generalizing and integrating geospatial data (Bédard and Han 2008). Such frequency may lead to numerous modifications of data meaning which may cause confusion for end-users and hence complicate the interpretation of geospatial datacubes.
- geospatial datacube structure promotes a rapid and easy use of data. The rapidity and ease of data use may lead users to 1) misunderstand the inherent characteristics of data, and to 2) mistakenly feel that data are made-to-order for their decision analysis needs, and hence to deter them from adopting an informed behaviour towards data (Levesque et al. 2007).

In reality, during the interoperability process, each agent directly or indirectly deals with the risks of data misinterpretation by making decisions. Examples of decisions can be to ask the source agent for more information, to seek the help of domain experts to oversee data interpretation, etc. In order to make such decisions, agents need relevant information

about data as well as about the context in which data has been defined. In this thesis work, we intend to provide agents with such relevant information.

5.2.3 Causes of the risks of data misinterpretation

The main causes of data misinterpretation are the **imperfect resolution of the semantic heterogeneity** and the **semantic uncertainty**, which is mainly caused by the inappropriateness of production context with regards to the interoperability context. An example of context inappropriateness occurs when the resolution of remote sensing data of road network is not suitable to the resolution required by end-users. Context inappropriateness may occur also when the referencing systems associated with data (e.g., North American Datum of 1927 (NAD 27)) is different from the one used by end-users (e.g., North American Datum of 1983 (NAD 83)).

The risks of data misinterpretation may also be caused by the following issues:

- *Inherent imperfection of the communication process.* In order for the communication process to work properly, the destination should interpret data as close as it was originally intended, and adapt data to his/her/its context. For that, both source and destination need to have a common set of knowledge and beliefs (Schramm 1971). However, this is not always the case since the receiver may have a different understanding of data. In addition, during the communication process, a possible source of noise can affect the transmission of exchanged messages between agents. Consequently, the original intended meaning may be modified.
- *Vagueness of data meaning.* A vague representation may indicate a concept for which there are some objects that cannot be determined either to “fall under” the concept or not to fall under it (Kearns 1997). Vague geospatial data is hard to communicate among humans and harder among software agents (Cai et al. 2003).

The risks of data misinterpretation may occur each time data are transferred from an agent to another. That is, the risks may propagate along the chain of stakeholders in the interoperability process (e.g., software agents, human/software mediator, end-user, and lawyer). Risk propagation becomes more complicated with the increase of the number of

stakeholders in the process of the interoperability and may raise suspicion about the convenience of re-using data (the main aim of interoperability).

In the following section, we propose a brief review of the approaches dealing with risk of data misinterpretation when simultaneously using different information systems.

5.2.4. Managing the risks of data misinterpretation: an overview of approaches

These approaches vary from taking precautions that must be applied to avoid possible risks of data misinterpretation (i.e., prevention approaches) to more complex methods that propose to assist agents during data use (i.e., creative approaches).

5.2.4.1 Prevention approaches

Some prevention approaches (also called *a priori approaches*) define measures to prevent certain predicted risks of data misinterpretation. These measures vary from simple restriction of data access for certain individuals or groups, to more complex procedures such as training users in order to improve their ability to interpret data, enhancing data selection tools, providing metadata, and developing context-aware systems that help users to better adapt data to their specific use. An example of context-aware systems is NAMA (Kwon 2005). This system is based on a context-aware agent that predicts the behaviour of end-users and helps them to interpret the content of commercial Web sites.

However, a major limit of prevention approaches is that it is not easy (nor convenient) to predict all possible situations where the risk of misinterpretation can occur.

5.2.4.2 Reactive approaches

Reactive approaches (also called *a posteriori approaches*) propose to take certain actions while using data in order to reduce the risk of data misinterpretation. Some approaches communicate relevant information about data use or provide warnings to end-users when a risky operation is performed (e.g., measuring a distance without having the map units defined) (Beard 1989, Hunter 2000, Levesque et al. 2007) or when a pattern is not

considered. Other approaches propose to figure out the intended meaning of data when using different information system (i.e., interpreting data on the fly).

The approach proposed in this thesis is comparable to reactive approaches. The approach, which will be detailed in sections 5.3 and 5.4, aims to define and assess the risks of data misinterpretation by evaluating -on the fly- the fitness-for-use of datacube schemas and of production contexts with regards to the context of the interoperability (e.g., the purpose of interoperability and the specific requirements of end-users).

5.3 An approach to identify and evaluate the risks of misinterpretation in semantic interoperability of geospatial datacubes

We aim to deal with the risk related to data misinterpretation during the interoperability between geospatial datacubes in a systematic manner based on the risk management paradigm. This paradigm is well-known for identifying potential risks, evaluating them, and making decisions about them in project management (Morgan 1990, Renn 1998, Afla and Smith 2007). This paradigm appears to be a rich framework that can help software agents and human stakeholders to identify the risks in a systematic manner, and make appropriate decisions about these risks during the interoperability of geospatial datacubes.

Accordingly, we define three iterative phases to manage the risk related to data misinterpretation: 1) identifying the risks related to data misinterpretation, 2) evaluating the risks, and 3) reacting to the risks by making appropriate decisions related to the semantic interoperability process. Possible reactions to the risks include: reduction (i.e., reducing the risk to an acceptable level), absorption (i.e., enduring the consequences of the risk), and transfer (i.e., shifting the risk from one entity to another).

Identifying and assessing the risks are essential steps to determine how to manage them. They can be conducted simultaneously (Webster et al. 2005, Sester 2000). In order to identify and evaluate the risks of data misinterpretation in the interoperability between geospatial datacubes, we measure the fitness-for-use of the conceptual models of geospatial datacubes with regard to the requirements of end-users. We remind that a conceptual model

of a geospatial datacube includes the elements of both schema (e.g., geometric primitives, the number of levels, the order of levels, and type of hierarchy) and production context (e.g., method used to summarize data, referencing system and geospatial coverage).

5.3.1 Fitness-for-use of geospatial datacube conceptual models to identify and evaluate the risks

Conceptual models are central to information system design, and are used as a basis for developing information systems to meet as close as possible user's requirements at different levels (e.g., application, enterprise and industry levels) (Wand and Weber 2002, Moody 2005).

However, the quality of conceptual models may be insufficient for certain data reuse. Poor fitness-for-use of conceptual models may cause a risk of misinterpreting data, and may undermine the reuse of geospatial data, i.e., the main aim of interoperability. In fact, during data interpretation, agents make decisions when assigning a meaning to the received data. Examples of decisions include considering that a geospatial object has a broad boundary (e.g., a partial cut for a forest stand), although it is represented using a crisp polygon, and considering that two geospatial dimensions are more or less similar, although they have different hierarchies. Poor fitness-for-use of conceptual models makes agents uncertain about such decisions and, hence, affects the interpretation process and increases the risks of misinterpreting data. On the other hand, a good fitness-for-use of conceptual models makes agents comfortable in making decisions during data interpretation and, hence, decreases the risks of misinterpreting data. Consequently, evaluating the fitness-for-use of conceptual models can help agents in evaluating the risks and hence in making appropriate decisions related to such risks. For example, based on such fitness-for-use agent can be advised:

- a. not to use one of the heterogeneous elements of the conceptual models (the one which has a lower fitness-for-use compared to the other).
- b. to consider a model element which has an excellent fitness-for-use.
- c. to use a model element which has a good fitness-for-use to create new one.

- d. not to consider two elements of two conceptual models if they both have a poor fitness-for-use that is likely to produce a poor result.

The subject of the evaluation of conceptual models quality has occupied a substantial part of the effort devoted to conceptual modeling (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 (Atkinson and Kühne 2003, Genero et al. 2007). However, research works focused more on schema quality than the quality of the production context. Moreover, while a range of quality frameworks have been proposed in the literature, none of these has been widely accepted in practice and none has been considered as a potential standard (Moody 2005). As a result, conceptual models have been evaluated in an *ad hoc* manner (Moody 2005).

5.3.1.1 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” or a “bad” schemas, there are neither guidelines nor standards for evaluating the quality of conceptual schema (Moody 2005). A number of quality criteria have been proposed in the literature, but none of these has been widely adopted and none has been considered for standardisation. Typical criteria are completeness, understandability, and minimality (Akoka et al. 2007):

- Completeness: a conceptual schema is complete when it contains all needed features of the application or of the domain (Batini et al. 1992). The completeness can be measured by the degree of which the conceptual schema covers users’ requirements (Akoka et al. 2007).
- Understandability: it refers to the easiness with which users interpret schema elements. It depends on how much schema elements are made explicit (Akoka et al. 2007).
- Minimality: a schema is said to be minimal when every elements of the requirements appears only once (Batini et al. 1992).

Due to the absence of any consensus about how quality should be evaluated, people continue to evaluate conceptual schemas in an *ad hoc* and subjective manner based on common sense and experience (Moody 2005).

5.3.1.2 The quality of production context

In chapter 2, we distinguished between two kinds of context: production context which specified by data producer, and use context which refers to the characteristics that surround user's application. Production context is described in a way to be suitable for the intended use of data. However, a production context which is appropriate for a given use may be less appropriate for another. That is, a production context which is suitable for the purpose for which a datacube has been collected may be less appropriate for the purpose of the interoperability (e.g., the application for which the interoperability is carried out). For example, the dates of certain photographs displayed on Google Earth are often one or more years old, which may have no impact on several usages but may mislead others.

Experiences showed that production context has several limitations and is rarely consulted by end-users (Timpf et al. 1996, Frank 1998, Harvey et al. 1999, Devillers et al. 2002). In fact, production context is usually defined using technical descriptions which are often difficult to understand by non-experts users. Moreover, context may be ignored or ill-described. This raises questions about the appropriateness of production context for a given application. Such appropriateness is indicated by the external quality of context (its fitness for the application purpose). A poor quality of context indicates a higher risk, and may undermine the reuse of geospatial data (Agumya and Hunter 2002), i.e., the main aim of semantic interoperability. On the other hand, a good quality of context makes it easier to understand data within that context and, hence, decreases the risks of misinterpreting data. Consequently evaluating the quality of context would help to identify and assess the risks of data misinterpretation. The worst the quality of context information is, the higher the risk of data misinterpretation is.

While, significant research efforts have been carried out to evaluate and enhance the quality of geospatial data (e.g., Agumya and Hunter 2002, Frank et al. 2004, Devillers et al. 2007, Frank 2007, Boin 2008), there has been no work to evaluate the fitness-for-use of context

information associated with geospatial data. We should notice, however, that many works have proposed ways to identify and represent context information and, hence, enhance their fitness-for-use. A typical example of these works is the ISO/TC 211 19115:2003 Geographic information - Metadata standard (ISO/TC 211, 2003), which defines a common set of metadata terminology, definitions and extension procedures. However, as it is discussed in chapter 2, semantic interoperability goes beyond standardizing techniques and takes into account the diversity of data and context representations. Our approach of semantic interoperability between geospatial datacubes takes into account the diversity of context representations. It tries to find a way to manage the risks of misinterpreting geospatial datacubes content despite the difference in quality of their contexts.

In order to evaluate the fitness-for-use of production context and conceptual schema, we define a set of indicators that helps agents (software agents or human stakeholders) to reason about context and make appropriate decisions about data interpretation. These indicators show the relevance of geospatial datacube's conceptual model for the purpose of interoperability (i.e., the purpose for which the interoperability is carried out).

5.3.2 Indicators to identify and evaluate the risks in the semantic interoperability between geospatial datacubes

In this section, we propose a restricted set of indicators to evaluate the fitness-for-use of production context and the one of geospatial datacube schema. We also propose a method to evaluate these indicators.

We define two categories of indicators: a category containing those which are related to the schema (*relevance of the geometric primitive*, *relevance of the structure* and *relevance of the hierarchy order*) and another category which contains those related to the production context (*relevance of production context*, *freshness of production context* and *trust of production context*) (c.f., 5.3.2.1 and 5.3.2.2). Each indicator is evaluated according to a function. The resulting quality is depicted by a value within the interval (0, 1). The value 1 indicates perfect quality, and hence a minimum risk. The value 0 indicates completely poor quality, and hence a maximum risk. This value is defined in a pragmatic and mathematical way as it is 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 superior or equal ($= >$), inferior ($<$) to compare the fitness-for-use of elements of different production contexts or different schemas. In other words, we can say that a quality 0,8 is sufficiently better than a quality 0,4, but we cannot say that it is precisely twice better. Also, based on this ordinal measure, we assign a qualitative value (i.e., “good”, “medium” or “poor”) to the fitness-for-use of both production context and of schema. In using such qualitative scale of measures, we make it easier for stakeholders to understand the resulted values of the fitness-for-use.

We should notice that the proposed indicators do not aim at being complete or precise but rather at making agents aware of the fitness-for-use of context information and of schema.

In our method, in order to simplify risk evaluation, we suppose that the requirements of end-users are represented in the form of a geospatial datacube model (using the same paradigm as for the models of the datacubes to be interoperated). The end-users would like to analyze the risk of forest fire on population according to the region, time, forest stand and age category. Accordingly, we add the conceptual model shown in Figure 5.2, which describes the requirements of end-users, to our example introduced in chapter 3 (see Figure 3.8).

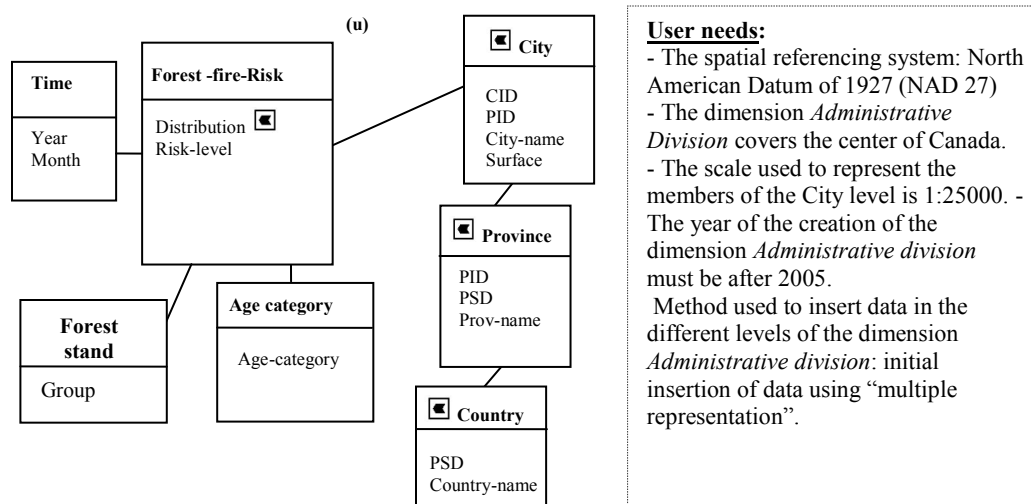


Figure 5.2: Geospatial datacube model describing the needs of the end-user.

5.3.2.1. Indicators for evaluating the external quality of geospatial datacube schema

In order to evaluate the quality of the schema, we define three indicators: *relevance of the geometric primitive*, *relevance of the structure* and *relevance of the hierarchy order*.

1- Relevance of the geometric primitive: this indicator (denoted G_s) evaluates the relevance of the source geometric primitive (i.e., a primitive used to represent an element of the model of a datacube to be involved in the interoperability) compared to the required geometric primitive (i.e., a primitive used to represent an element of the datacube model describing the needs of end-users). That is, the value of this indicator depends on how end-users would prefer geospatial members, which would be resulted from the interoperability process, to be represented (simplified representation, and faithful representation, etc.). For example, the relevance of the geometric primitive can be evaluated based on the degree of faithfulness with which the source geometric primitive can be considered as a cartographic generalization of the required geometric primitive. Table 5.1 represents an example for evaluating the relevance of the geometric primitive according to the degree of faithfulness. The values of this table are defined taking into account the rules of cartographic generalization and the common practice of cartographers. In this example, we set the values 0, 0.25, 0.5, 0.75, and 1 respectively when the source geometric primitive:

1. cannot be considered as a cartographic generalization of the required geometric primitive (e.g., line (1D¹⁵) cannot be considered as a generalization of point (0D)),
2. cannot be considered as a cartographic generalization of the required geometric primitive, however, the result of generalization will be a bad representation of reality (e.g., a point (0D) and a line (1D)),
3. cannot be considered as a cartographic generalization of the required geometric primitive, however, the result of generalization will not faithfully reflect reality (e.g., 2D and 3D),
4. cannot be considered as a cartographic generalization of the required geometric primitive, this result will faithfully reflect reality (e.g., 0D and 3D).),
5. is similar to the required geometric primitive.

¹⁵ D indicates Dimension (e.g., 0D is 0 Dimension).

Geometric primitives are defined for the levels or measures of a geospatial datacube. Consequently, this indicator is evaluated for measure and level layers defined in the proposed framework of the communication between datacubes agents (see Figure 4.2).

		Required geometric primitive			
		0D	1D	2D	3D
Source geometric primitive	0D	1	0.25	0.5	0.5
	1D	0	1	0.75	0.5
	2D	0.25	0	1	<u>0.75</u>
	3D	0.5	0	0.75	1

Table 5.1: Evaluating 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 a required element. It indicates the ratio of the element's components which are semantically related to those of the required element. As such, the relevance of a dimension structure is evaluated based on the ratio of the *number of hierarchies*. Similarly, the relevance of a hierarchy is evaluated based on the ratio of the *number of levels*. Also, the relevance of a level is evaluated according to the ratio of the *number of attributes*. The relevance of the structure P_S is calculated using the formula (1):

$$P_s = \begin{cases} \frac{N_{CE}}{N_{CR}} & ; \text{if } N_{CE} < N_{CR} \\ 1 & ; \text{Otherwise} \end{cases} \quad (1)$$

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

This indicator can be evaluated for all layers of our framework of interoperability between geospatial datacubes (see Figure 4.2).

3- Relevance of the hierarchy order: it indicates the relevance of the order of the datacube hierarchy with regard to the order of the required hierarchy. Let us 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 the 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:

$$o_s : \forall n_1, n_2 \in H, \forall n_1', n_2' \in B : [n_1 < n_2] \Rightarrow n_1' < n_2' \quad (2a)$$

Where n_1 and n_1' are two semantically related levels. Idem for n_2 and n_2' . The expression $n_1 < n_2$ (or $n_1' < n_2'$) is true means that n_1 rolls-up to n_2 (or n_1' rolls-up to n_2'). The value 1 (or 0) is set to the elementary relevance o_s when the expression (2a) is true (or false).

$$O_s = \sum o_s / No_s \quad (2b)$$

Where No_s is the number of the elementary relevance o_s .

As its name indicates, the indicator *relevance of the hierarchy order* is evaluated only for the hierarchy layer defined in the proposed communication framework.

5.3.2.2 Indicators for evaluating the fitness-for-use of production context

In this section, we propose a set of indicators and a quantitative approach to evaluate the fitness-for-use of production context. The fitness-for-use of production context refers to the relevance of information about this context for a specific use. For example, a high precision of geospatial data can be ranked very low in terms of degree of relevance for touristic use where priority is given to semantic map's content. That is, the type of touristic content does not necessitate a high geometric precision. On the other hand, the precision may be very relevant when a map is generated for cadastral use. The proposed indicators are: *relevance of production context*, *freshness of production context* and *trust of production context*.

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

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

Where N_{Elt} is the number of thematic, geospatial, or temporal elements of production context which are semantically related to those required by end-users, N_{ReqElm} is the total number of required production context elements, and w a predefined value between 0 and 1 which indicates the level of importance of each type (i.e., thematic, geospatial, and temporal) for end-users. If N_{Elt} is equal or larger than N_{ReqElm} , the production context is perfectly relevant, and thus has 1 as value.

2- Freshness of production context. This indicator indicates the degree of freshness of the production context of the geospatial datacubes with regards to the requirements of end-users. It is evaluated according to the age of production context with regard to their lifetime. The age of production context is the time passed since the date of the definition of production context (T_{def}) until the desired date of freshness of this production context (T_{req}). The lifetime is the number of years, after which the production context will not be valid anymore. The freshness of production context A_m is evaluated based on the following formula:

$$A_m = \begin{cases} 1 - \frac{|T_{req} - T_{def}|}{DV} & ; \text{ if } |T_{req} - T_{def}| < DV \\ 0 & ; \text{ Otherwise} \end{cases} \quad (4)$$

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

3- Trust of production context. This indicator describes the degree of faith that we have in the provided production context. A decrease of trust lessens the fitness-for-use of context information. Generally, in a chain of stakeholders, if information is transmitted in a sequential manner (i.e., transmitted once to one stakeholder after another), the trust decreases with the number of stakeholders (Bédard 1986, Moe and Smite 2007). We evaluate the trust using the following formula:

$$T_m = \frac{\sum \alpha_i}{N} \quad (5)$$

Where α_i is the confidence given to the i^{th} stakeholder. The value of confidence is between 0 and 1. N is the number of stakeholders that transmitted the production context element. We consider that each stakeholder transmits a production context element just once.

We should remind here that the set of above indicators of geospatial datacube model (schema and production context) does not aim at being exhaustive or precise but rather at making stakeholders aware of the risks of data misinterpretation, and helping them to make reasoned decisions in dealing with these risks during the interoperability between geospatial datacubes. Such a method is frequently used in several fields which involve factors that are difficult, even impossible, to evaluate in an exhaustive and precise way (e.g., the fields of epidemiology, ecology, economic sciences). Such a method was used to evaluate the quality of geospatial data by Devillers et al. (2007).

The way of presenting these indicators to the user has a great importance on decision-making. Generally, the decision-support systems use a restricted number of indicators (Few 2006). In this regard, we suggest to present only two indicators to the user: one to indicate the fitness-for-use of schema (denoted Q_s) and the other to indicate the fitness-for-use of production context (denoted Q_m). If the user would like to get more information about the

fitness-for-use of schema or of production context, he/she can go into each of the defined indicators. The two overall indicators Q_s and Q_m are calculated as follows:

$$Q_s = \frac{\sum (a_i \times I_{si})}{n} \quad (6)$$

$$Q_m = \frac{\sum (b_j \times I_{mj})}{m} \quad (7)$$

Where I_{si} and I_{mj} refer, respectively, to the value of the i^{th} indicator of schema and the value of j^{th} indicator of production context. The variables n and m are the numbers of the indicators. a_i and b_j are predefined values between 0 and 1 that indicate the importance of each quality indicator of schema and production context, respectively.

In the following section, we propose a general framework to support stakeholders (software agents or potential human stakeholders) in responding to the risk of data misinterpretation. The framework suggests presenting the previously defined indicators to stakeholders in an intuitive way to help them making decisions about such risks. In section 5.5, we show an example of how the proposed indicators can help making appropriate decisions about the risks of data misinterpretation.

5.4 Responding to the risks of data misinterpretation in the semantic interoperability of geospatial datacubes

The aim of this section is to propose an approach to help software agents or potential human stakeholders to make appropriate decisions about the risk of data misinterpretation in the context of the interoperability between geospatial datacubes. The approach consists of a general framework and an algorithm. The proposed framework takes into account the proposed categorization of semantic heterogeneity. Also, it corresponds to the conceptual framework of interoperability proposed in chapter 4. It suggests responding to the risks at one level at a time (from the more general level to the most detailed level). The algorithm shows how the framework can be implemented to support software agents.

Responding to the risk may involve reducing, absorbing, or transferring such a risk.

- *Risk reduction.* It consists of reducing the risk to an acceptable level. The reduction of the risks of data misinterpretation can be achieved by 1) resolving the context incompatibilities that may occur in different dataset, and 2) adding new context information to decrease the uncertainty in geospatial data. However, geospatial context involves a large number of elements (e.g., geospatial position, form). Thus, for practical reasons, risk reduction should target elements where the maximum risk reduction can be achieved (Agumya and Hunter 2002). Accordingly, the reduction of risk should consider only context elements to which a particular decision is most sensitive, or where vulnerability to the consequences of data misinterpretation is highest. Although the risks of misinterpreting geospatial data can be reduced, it cannot be completely eliminated (Bédard 1988). There is always a need for risk absorption.
- *Risk absorption.* It is the most commonly used approach (Bédard 1986, Agumya and Hunter 2002) and consists of enduring the consequences of risks. For data misinterpretation, absorbing risks consists of 1) assuming that the interpretation of geospatial data may not be perfectly appropriate but sufficient, and 2) being able to endure the consequence of misinterpreting data.
- *Risk transfer.* It consists of shifting the risk from one entity to another (Agumya and Hunter 2002). Risk transfer may require insurance contracts or policies to reimburse damage related to the risk of data misinterpretation (Gervais 2004). The risk of data misinterpretation can be transferred when a stakeholder (e.g., software agent) is not able to make a decision about the risk. Then, he/she could transfer the risk to another stakeholder (e.g., human stakeholder).

Based on these well established possibilities of responding to the risks, we define a general framework to support the response to the risks of data misinterpretation in the semantic interoperability between geospatial data cubes.

5.4.1 General framework to respond to the risks of data misinterpretation

We propose a general framework that aims at guiding stakeholders 1) to identify the risks of data misinterpretation based on the previously identified indicators, and 2) to make decisions to deal with such risks. It consists of five successive phases of analyzing and responding to the risks of data misinterpretation. These phases correspond to the five layers of our conceptual framework of the interoperability between geospatial datacubes (c.f., chapter 4), from the more general layer to the most detailed: cube, measures, dimensions, hierarchies, and levels. At each phase, the stakeholder analyzes the quality of schema and production context based on the previously identified indicators, and make some decisions about the risks of data misinterpretation. Such decisions include (see Figure 5.3):

- To suspend the interoperability process of the geospatial datacubes if data misinterpretation presents a high risk that can lead to harmful consequences. In this case, the stakeholder does not need to continue with the remaining phases.
- To continue the interoperability process of the geospatial datacubes if there is no risk of data misinterpretation (e.g., the model elements are not semantically related) or if data misinterpretation does not present a high risk. Two decisions can then be made:
 - Solving the causes of the risks. For example, to solve the problems of difference between geospatial referencing systems by choosing a common referencing system.
 - Doing nothing to solve the causes of the risks, and enduring the risks if they do not significantly affect data use. For example, a difference in accuracy of 1 meter for geospatial data does not affect a tourist application, so the stakeholder can endure this heterogeneity.

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

1. First, to help stakeholders to identify the risks of data misinterpretation. In fact, at each phase of the proposed framework, while a good quality of schema and context indicates it is less likely to have a risk of data misinterpretation, a poor quality indicates a higher risk.
2. Second, to help stakeholders to make appropriate decisions at each phase (to suspend the interoperability process, to solve or endure the problems). For example, if two heterogeneous elements, which are essential for the interoperability, have a poor quality of context, stakeholders are advised not to consider both elements in any interoperability result (e.g., common model).
3. Third, to help stakeholders to solve the semantic problems at each phase of the proposed framework. The indicators are especially useful in the situations where there are choices to be made in connection with the semantic uncertainty (c.f., section 4.4). In fact, based on these indicators, stakeholders can be advised to consider an element which has an excellent fitness-for-use, or not to consider two heterogeneous elements if they both have a poor fitness-for-use.

At the end of each phase, according to the decision made (to suspend the interoperability process, to solve the causes or to endure the risks, etc.), the stakeholder should write a report explaining (1) the reasons of the suspension, (2) how the problem was solved and noted comments, or (3) 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:

- The approach allows stakeholders to make relevant decisions about the risks of data misinterpretation at an early stage of the interoperability process. This allows them to put less time and effort dealing with the problem of heterogeneity and other causes of the risks. Indeed, at each phase, stakeholders can suspend the interoperability before going into details and hence reduce the costs of analysis and integration. Stakeholders can also continue the

interoperability by taking into account the observations made at one phase to better deal with the risks at a later phase.

- The proposed approach is in accordance with the mental model of human (Yougworth 1995, Rivest et al. 2005, Bédard and Han 2008). Indeed, this framework is based on a hierarchical structure which is one of the essential principles of human cognition (Edwards 2001). This principle stipulates that humans gather data and context information 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 reuse of data with the minimum effort (Rosch 1978).

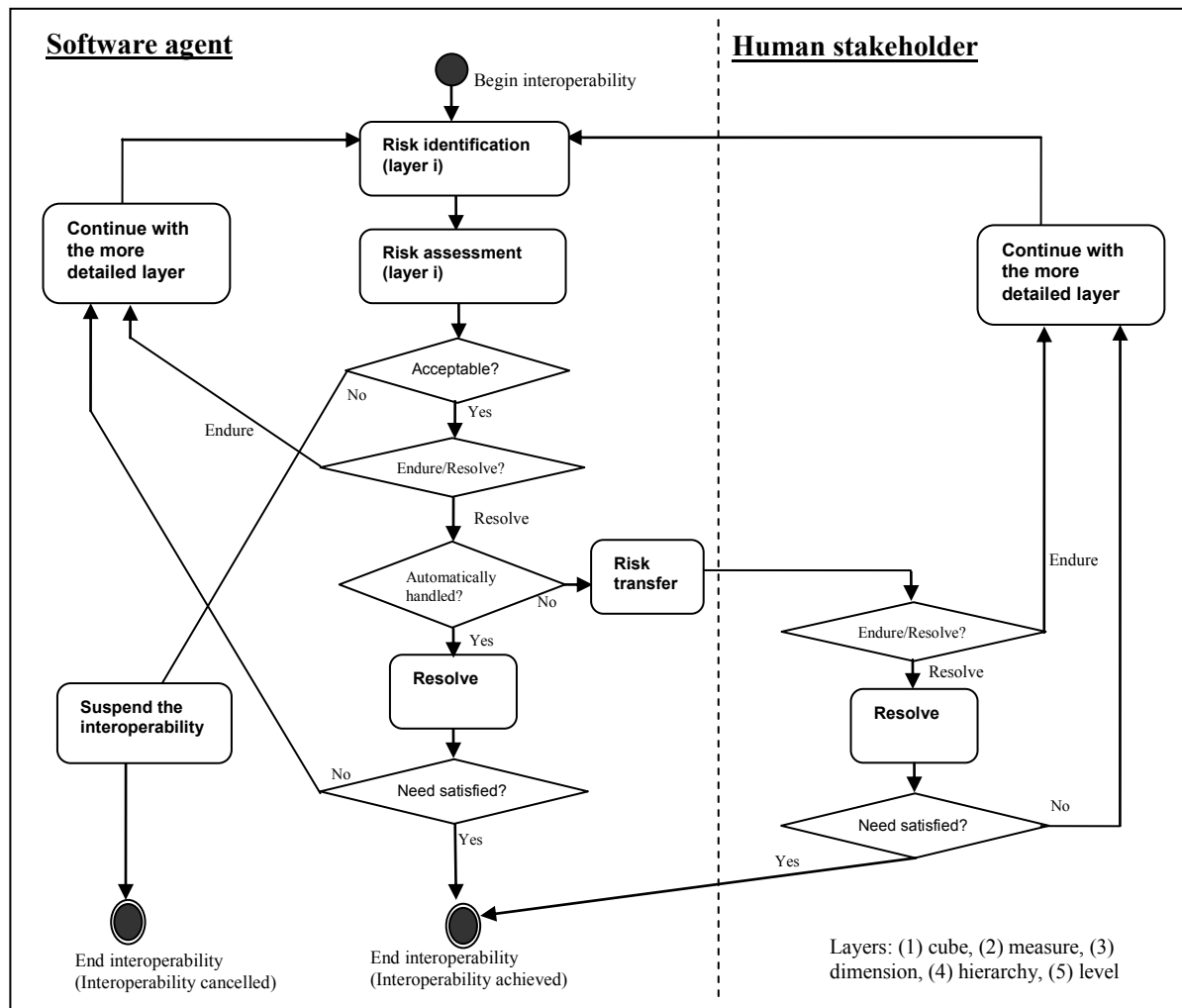


Figure 5.3: A general framework to respond to the risks of data misinterpretation.

Figure 5.3 illustrates the proposed framework. Based on the previously identified indicators of the fitness-for-use of schema and production context, stakeholders start by identifying the risks of data misinterpretation at the more general layer (i.e., cube layer) and making decisions to solve their causes (e.g., irrelevance of production context information). For example, identifying dissimilar geospatial referencing systems, one can propose to use a common referencing system for the heterogeneous datacubes. Then, stakeholders identify the risks at the measure layer and respond to them. For instance, they can identify a potential conflict related to the difference in the aggregation methods (e.g., aggregating using the function geometric union versus the function centre of gravity), and decide to use only one of these methods. Then, stakeholders identify the risks at the dimension layer and make decisions to solve these risks. For example, they can decide to suspend the interoperability between geospatial datacubes if there is a conflict related to the difference of the coverage of geospatial data (e.g., the province of Quebec versus the province of Manitoba). After analyzing the risks at the dimension layer, stakeholders identify the risks at the hierarchy layer and make decisions to solve them. For example, they can decide to suspend the interoperability when two semantically related hierarchies belonging to different geospatial datacubes have incompatible levels. Finally, stakeholders identify the risks related to the level layer, and makes decisions about these risks. For example, one can decide to ignore the heterogeneity of scale of cartographic representation used for the members of two levels belonging to different geospatial datacubes (e.g., 1:125000 and 1:120000). Later on, we provide a continuation of our example presented in Figure 3.8 to show how the proposed framework can help to deal with the risks of data misinterpretation.

The proposed framework can be implemented using an algorithm that guides the stakeholder in making sound decisions when dealing with the risks of data misinterpretation. The algorithm defines a threshold under which the risk of data misinterpretation is unacceptable.

5.4.2 An Algorithm to support decision-making in the semantic interoperability of geospatial datacubes

In the following algorithm, OFuEl is the overall fitness-for-use of each element of a datacube model (i.e., measure, dimension, hierarchy, or level).

Begin

Input: - c_1, c_2 are two concepts of the models.

- *maq* (minimum accepted fitness-for-use) denotes a threshold under which the fitness-for-use is likely to cause an interoperability failure.

For each couple of concepts (where semantic similarity $(c_1, c_2) > 0$)

 Compute (OFuEl (c_1) and OFuEl (c_2))

If (OFuEl (c_1) < *maq* and OFuEl (c_2) < *maq*), then:

 Display message: “Be careful, both concepts have a very poor fitness-for-use. It might be better not to consider both concepts”

Else if (OFuEl (c_1) < *maq*), then:

 Display message: “Be careful, c_1 has a very poor fitness-for-use, you should not choose c_1 ”

Else if (OFuEl (c_2) < *maq*), then:

 Display message: “Be careful, c_2 has a very poor fitness-for-use, you should not choose c_2 ”

Else If (OFuEl (c_1) > OFuEl (c_2))

 Display message: “ c_1 has a better fitness-for-use, you better choose c_2 ”

Else Display message: “ c_2 has a better fitness-for-use, you better choose c_1 ”

End If

End For

End

Along with the content of the resulting message, we define a set of warning symbols indicating the degree of the risk of data misinterpretation during the interoperability process. This allows stakeholders to intuitively understand the message.

5.4.3 Symbolic notation of quality indicators

We provide a set of warnings which are based on standard danger symbols proposed by ANSI Z535.4 (1991) and previously used in the geospatial domain by Levesque et al. (2007). These symbols are enriched by others to suitably assist stakeholders in geospatial datacube interoperability. The symbols show the degree of the risks of data misinterpretation at each layer of the proposed conceptual framework, and thereby stimulate appropriate responses to such risks. For example, if the warning is ‘Danger’, it would be better to suspend the interoperability process which may lead to undesirable consequences. Table 5.2 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 quality indicators	Significance of the symbol
$Q = 0$		Stakeholders are advised to stop the process of interoperability.
$0 < Q \leq 0.25$		There is a high risk if the stakeholders decide to continue the interoperability.
$0.25 < Q \leq 0.5$		Stakeholders are warned about the existence of potential risks.
$0.5 < Q \leq 0.75$		Stakeholders are invited to pay attention if they decide to continue the interoperability.
$0.75 < Q < 1$		Information will be shown to stakeholders.
$Q = 1$		Stakeholders are informed that the quality is very good, and are invited to continue with a more detailed level (if there is any).

Table 5.2: A definition of the symbols according to the quality value.

To the above list, we add: (1) the symbol  which invites stakeholders 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 context information), 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 the risks of data misinterpretation.


5.5. Example of application

We continue with our example (Figure 3.8 showing two geospatial datacubes and Figure 5.2 showing the requirements of end-user in the form of a geospatial datacube model). 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 the involved geospatial datacubes.

To simplify the reading of this section, we evaluate the indicators at only three layers of our conceptual framework (i.e., cube, measure and dimension)¹⁶. Also, we set the value 1 for a and b (which indicate the importance of each quality indicator of schema and production context respectively; see equations (6) and (7)). We should note that, in this example, we do not consider other elements of production context than those described in the metadata. Consequently production context is indeed metadata.


1. Cube layer:

When two datacubes source are semantically related (e.g., by their objective, their name or their contextualized measures, i.e., by taking into account their dimensions), then their schema and context are analyzed, their quality indicators are evaluated and a decision is made about the continuation of the interoperability process.


In our example (c.f. Figure 3.8), the datacubes do not have a common subject of analysis. Accordingly, there is no semantic conflict between the geospatial datacubes. Then, the symbol  will be shown.

2. Measure layer:

When two measures are semantically related (e.g., by their name, their definition, or their descriptive or geometric representation), then their schema and their production context are analyzed, the quality indicators are evaluated and a decision is made about the continuation of the semantic interoperability.

In our example, the measures of both geospatial datacubes have not been defined for the same subject of analysis. Accordingly, there is no semantic conflict at the measure layer. The stakeholder is invited to continue to the next layer (dimension layer). Then, the symbol  is shown.

3. Dimension layer:

If the dimensions of the two datacubes are not semantically related, the stakeholder does not have to go into more details for these dimensions, the symbol  will be shown, and the stakeholder will be invited to continue with other possible combinations of dimensions. It is the case for dimensions *Age category* of datacube *C1* and *Region* of datacube *C2*.

If there is a semantic relation between two dimensions, such as the dimensions *Administrative region* of datacube *C1* and *Region* of datacube *C2*, then the quality of both dimensions is evaluated for their schema and their production context.

a. *Fitness-for-use of schema:*

➤ *Relevance of the number of hierarchies (relevance of structure)*

The dimension *Administrative region* of the datacube *C1* has only one hierarchy: (*Country*, *Province*, *Territory* and *City*), whereas the dimension *Region* of the datacube *C2* has two hierarchies: (*Country*, *Province*, *Territory*, and *City*) and (*Country*, *State*, and *City*). We remind that end-users need only one hierarchy: (*Country*, *Province*, and *City*). Then, the quality indicator *Relevance of the number of hierarchies* is evaluated according to the formula (1):

¹⁶ A complete version of this example is available in Annex D.

For $C1$, $P_s(\text{Administrative region}, \text{Administrative division}) = 1/1 = 1$


For $C2$, since the number of hierarchies of dimension *Region* is larger than that of the semantically related dimension of the required datacube $C3$ (*Administrative division*), then:

$$P_s(\text{Region}, \text{Administrative division}) = 1$$

According to the formula (6) the quality of datacubes schemas:

$$Q_s(\text{Administrative region}, \text{Administrative division}) = 1$$

$$Q_s(\text{Region}, \text{Administrative division}) = 1$$

Consequently, the symbol  will be shown for the two dimensions *Administrative region* and *region*. In the following phases (hierarchy and level), the analysis will be refined to lead to a choice among the dimensions of datacubes sources aiming to fit end-user requirements (expressed in the model C3).

b. Fitness-for-use of production context:

➤ *Relevance of production context*

Production context associated with the dimension *Administrative region* of $C1$ have two elements which are semantically related to the elements required by end-users (the spatial coverage and the year of creation of the dimension). On the other hand, production context associated with the dimension *Region* of $C2$ have only one element which is semantically related to an element required by end-users (the spatial coverage of the dimension). If we consider that the level of importance of spatial information is 1, then based on formula (3):

$$P_m(\text{Administrative region}, \text{Administrative division}) = 2/2 = 1$$

$$P_m(\text{Region}, \text{Administrative division}) = 1/2 = 0.5$$

➤ *Freshness of production context*

Production context of the dimensions *Administrative region* of *C1* and *Region* of *C2* were created respectively in 2002 and 1982. Moreover, these two dimensions have the same lifetime: 30 years. Consequently, according to formula (4):




$$A_m(\text{Administrative region}, \text{Administrative division}) = 1 - (2005 - 2002 / 30) = 0.9$$

$$A_m(\text{Region}, \text{Administrative division}) = 1 - (2005 - 1982 / 30) = 0.23$$

According to the formula (7) the quality of production context:

$$Q_m(\text{Administrative region}, \text{Administrative division}) = (1 + 0.9) / 2 = 0.95$$

$$Q_m(\text{Region}, \text{Administrative division}) = (0.5 + 0.23) / 2 = 0.36$$

Consequently, the symbol  will be shown for the dimension *Administrative region*. On the other hand, the symbol  will be shown for the dimension *Region*. The stakeholder is then invited to be careful when considering the dimension *Region* in the process of interoperability. Moreover, the stakeholder is invited to evaluate the quality of the detailed levels of these two dimensions (i.e., hierarchy and levels) by taking into account the difference in quality of these dimensions. We should note that, if the quality of one of two dimensions was very poor, then the symbol  would have been shown and the stakeholder would have been advised not to consider the detailed levels of this dimension.

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. It is the case of dimensions *Age category* and *Forest stand* in our example.

Figure D.1 shows an example of model which could be obtained to enable the interoperability between geospatial datacubes *C1* and *C2* according to the different phases of the general framework (see Figure 5.3) and using the proposed quality indicators. It is important to remind here that the proposed approach does not aim at finding a specific solution for the interoperability between geospatial datacubes, but rather sufficiently

informative to help stakeholders 1) to analyze the feasibility of the interoperability goal (e.g., define a common model) based on a top-down approach, and 2) to support the interoperability based on a set of indicators of the fitness-for-use of geospatial datacube schema and production context. These indicators help stakeholders to make appropriate decisions such as considering or not certain elements of the datacubes sources.

We should notice that the indicators and the method to identify and evaluate the fitness-for-of context can be useful in automatic context reasoning. We demonstrated such usefulness by proposing a way to implement the proposed indicators using existing semantic Web technologies such as OWL. This allows the notion of the fitness-for-use in semantic interoperability to be implemented and used on the Web. In fact, such technologies can describe the fitness-for-use indicators in a manner understandable by machines (software agents), and allow to define a number of rules to facilitate automatic context reasoning. More details about this work can be found in the paper titled “Modeling the External Quality of Context to Fine-tune Context Reasoning in Geospatial Interoperability” (Sboui et al. 2009b). A modified version of this paper is available in Annex C of this thesis.

5.6. Conclusions

The semantic interoperability between geospatial datacubes faces risks of data misinterpretation that may affect strategic decision-making process. Consequently, we need to deal with these risks by identifying them, assessing their severity, and making decisions about them. In this chapter, we proposed an approach to deal with the risks of data misinterpretation based on well established paradigm for managing the risks within project management. The approach consists of a set of indicators that allow to identify and evaluate the risks of data misinterpretation, and a general framework that aims to support stakeholders to make appropriate decisions about these risks. The proposed indicators evaluate the fitness-for-use of schema and production context (i.e., quality of context and of schema with regard to a specific use). They play an important role in managing the risks of data misinterpretation. They help stakeholders to identify the risks of data misinterpretation, to evaluate the severity of these risks, and to make appropriate decisions about them.

The proposed framework consists of five phases of analyzing and responding to the risks of data misinterpretation. These phases correspond to the five layers of our conceptual framework of the interoperability between geospatial datacubes, from the more general layer to the most detailed: cube, measures, dimensions, hierarchies, and levels. At each phase, the result of the evaluation of the proposed indicators is represented qualitatively using warning symbols. These symbols aim to make agents aware of the risks of data misinterpretation, and to help them making decisions in an intuitive way.

We should remind that the set of the proposed indicators and the proposed framework do not aim at being exhaustive or precise but rather at helping the stakeholder to make appropriate decisions to enhance the interoperability between geospatial datacubes.

We also presented an example of application that shows how the general framework as well as the proposed indicators can help stakeholders 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 between other information systems. For example, our approach can be implemented in existing tools to manage the risks related to the uncertainty about the match between geospatial concepts. In the next chapter, the proposed approach will be implemented to show how it enhances the semantic interoperability between geospatial datacubes.

Chapter 6: An experimentation of the proposed approach for the interoperability between geospatial datacubes

6.1. Introduction

As discussed, there is no tool that supports the interoperability between geospatial datacubes. Moreover, existing approaches of semantic interoperability are not very appropriate for geospatial datacubes. In addition, such approaches do not deal with the risks of data misinterpretation in a systematic way; consequently, the risks are still ill-defined and are ill-evaluated and, hence, are ill-managed. In chapters 4 and 5, we proposed a conceptual framework and an approach of managing the risks of data misinterpretation to support the interoperability of geospatial datacubes. In this chapter, we experiment the proposed conceptual framework and risk management approach using the example introduced in chapter 3 (see Figure 3.8). For that, we developed a prototype, called *Model for the Quality of Interoperability between Geospatial Datacubes (MQIGDC)*. The prototype uses the extended GsP (MGsP) to measure the semantic similarity between geospatial datacube concepts. The GsP was extended by embedding multidimensional properties (e.g., properties introduced in section 4.3.2). We remind that measuring the semantic similarity is out of scope of this thesis.

In the next section, we present the technology used to develop the MQIGDC prototype. In section 6.3, we present the architecture and the implementation of the MQIGDC prototype, and we test this prototype. Finally, we conclude this chapter in section 6.4.

6.2. Technology used

6.2.1 Software agents to mimic human communication

As discussed in chapter 2, human communication is viewed as an ideal analogy for the interoperability between information systems. Uitermark et al. (1999), Harvey (2002), Xhu

and Lee (2002), Brodeur et al. (2003), and Kuhn (2005) support this point of view. We base our approach of the interoperability on human-like communication processes.

Recently, software agent technology has played a key role in enabling communication between information systems (Nwana 1996, Payne et al. 2002, Brodeur 2004). According to (Nwana 1996), a software agent is defined as “a component of a software and/or hardware which is capable of acting exactly in order to accomplish tasks on behalf of its user”. Basically, software agents are components in an application that are characterized mainly by the ability to communicate (Wooldridge 2002). The ability to communicate means agents can interact with each others to achieve their goals (Zhao et al. 2008). More details about the description of software agents can be found in (Nwana 1996, Nwana and Wooldridge 1996). Software agents thus appear well suited to develop the interoperability between geospatial datacubes.

In chapter 4, we defined a communication framework which is based on software agents. Agents may be of two types: *datacube agents* and a *context agent*. Datacube agents represent geospatial datacubes to be involved in the interoperability process, and are engaged in a process of human-like communication. Datacube agents are defined according to five levels that correspond to the five conceptual levels of geospatial datacube; from the more general level to the most detailed level: cube, measure, dimension, hierarchy and level. Context agent's main role is twofold. First, it makes stakeholders (datacube agents or human stakeholders) aware of the context information related to the exchanged data. Second, it helps them to manage the risks of data misinterpretation. This is done by evaluating the fitness-for-use of schema and production context associated with geospatial datacubes.

In order to communicate, software agents must be able to understand each other although they use heterogeneous concepts (e.g., concepts having different meanings and represented similarly, or having the same meaning and represented differently). For that, they need to compare the semantics of the exchanged concepts and measure their semantic similarity. In our prototype development, we used the extended GsP tool to measure the semantic similarity between data to be exchanged between geospatial datacubes.

6.2.2 GsP tool to measure semantic similarity

Comparing the semantics of heterogeneous concepts belonging to different ontologies is usually done by reconciling two or more heterogeneous ontologies. This task is usually carried out by mapping, aligning or merging these ontologies (Giunchiglia and Yatskevich 2004).

In our work, in order to compare the semantics of geospatial data as well as the semantics of context¹⁷ associated with the exchanged data, we use the GsP tool which was implemented and tested in our research group. The GsP tool consists of comparing intrinsic and extrinsic properties of one concept against another. Intrinsic properties provide the meaning of a concept. Extrinsic properties refer to external factors (e.g., behaviours and relationships). The GsP tool has been extended (MGsP) by providing users the possibility to dig into multidimensional properties they need to focus on when comparing concepts.

In the next section, we present the details of the implementation and the experimentation of the MQIGDC prototype.

6.3. MQIGDC implementation and experimentation

While in our project we recognize the importance of ontologies, we argued that measuring the semantic similarity of geospatial concepts may require more than considering concepts meanings within different ontologies. Indeed, it may involve comparing other semantic aspects of the data to be exchanged, such as the fitness-for-use of data and the fitness-for-use of the context related to such data. These kinds of information are generally not included in ontologies. The MQIGDC prototype aims to show the feasibility of this idea.

In order to implement the MQIGDC prototype, we defined an architecture that takes into account the exchange of datacube concepts between software agents representing geospatial datacubes, and the decision-making support in the interoperability process.

¹⁷ In this chapter, we do not consider other elements of production context than those described in the metadata. Consequently production context is indeed metadata..

6.3.1 MQIGDC architecture

The architecture of the MQIGDC prototype contains three main modules: *Matching module*, *Fitness-for-use evaluation module*, and *Decision-making support module*, see Figure 6.1.

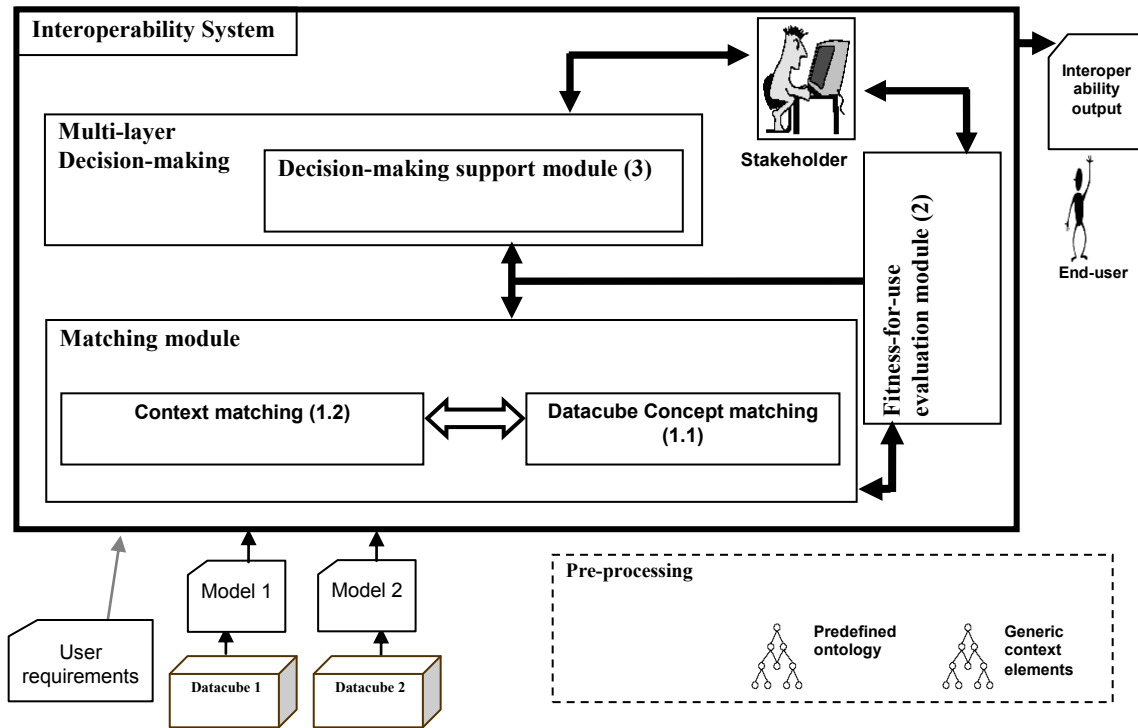


Figure 6.1: General architecture for geospatial datacube interoperability.

The modules are described as follows:

- *Matching module*. This module is based on the GsP tool (Brodeur 2004). In this module, geospatial datacube agents communicate together to exchange data between them. This communication is supported by a context agent who aims to make datacubes agents aware of the context associated with the exchanged data. The *matching module* aims principally at matching the concepts related to geospatial datacubes as well as those related to the contexts. In the *matching module*, datacube agents communicate using messages represented in XML documents. When an agent receives a message, it captures the concepts of the XML document and compares them

against the concepts stored in his/her/its ontology. More specifically, the module compares the received message's intrinsic (i.e., identification, attributes, attribute values, geometries, temporalities, and domain) and extrinsic properties (i.e., relationships and behaviours) with the intrinsic and extrinsic properties of a set concepts of the destination agent. The concepts are then sorted from the most similar to the least similar.

- *Fitness-for-use evaluation module*. It aims at evaluating the fitness-for-use of geospatial datacube models (i.e., schema and metadata). Such evaluation aims at identifying the risks related to the misinterpretation of datacube concepts (i.e., faulty interpretation or uncertainty about meaning).
- *Decision-making support module*. It aims at supporting users to make appropriate decisions about the risks of misinterpretation of datacubes concepts. This module analyzes the fitness-for-use of datacube schema and metadata with regard to the requirements of end-users. Then, it proposes some suggestions to help stakeholders better respond to the risks of misinterpretation. For example, if the fitness-for-use of the metadata associated with a received concept is very poor, the destination agent is invited not to consider the match of this concept with another one belonging to his/her/its ontology.

The previous three modules are interrelated. The *Fitness-for-use evaluation module* uses the output of the *Matching module* (i.e., a set of matched concepts to evaluate the fitness-for-use of their metadata and their schema). Also, the *Decision-making support module* uses the output of the *Fitness-for-use evaluation module* (i.e., the result of the evaluation of the fitness-for-use of datacube schema and metadata). Based on such result, suggestions will be made on how to deal with the risks of misinterpretation.

The modules use an ontology that contains the description of geospatial datacube concepts with their associated metadata. This ontology is implemented using Perceptory and manages the description of semantic, spatial, and temporal properties of datacubes concepts and their associated metadata. We should point out that, since no formal ontologies related

to the content of geospatial datacubes have been found, we defined the ontologies related to these datacubes based on their models.

In accordance with the GsP representation of geospatial concepts, we link the representation of the elements of geospatial datacube model (metadata element or schema element) with the abstraction class defined in GsP. Accordingly, a model element has a fitness-for-use which is characterised by a set of indicators (relevance of the geometric primitive, relevance of the structure and relevance of the hierarchy order, relevance of production context, freshness of production context and trust of production context). An indicator has a value (very good, good, medium, poor or very poor). As such, the representations of model elements and their fitness-for-use are embedded within the representation of geospatial datacube concepts (see the UML class diagram in Figure 6.2). Such embedment allows to use the extended GsP to measure the semantic similarity between elements of geospatial datacubes schemas and between elements of metadata.

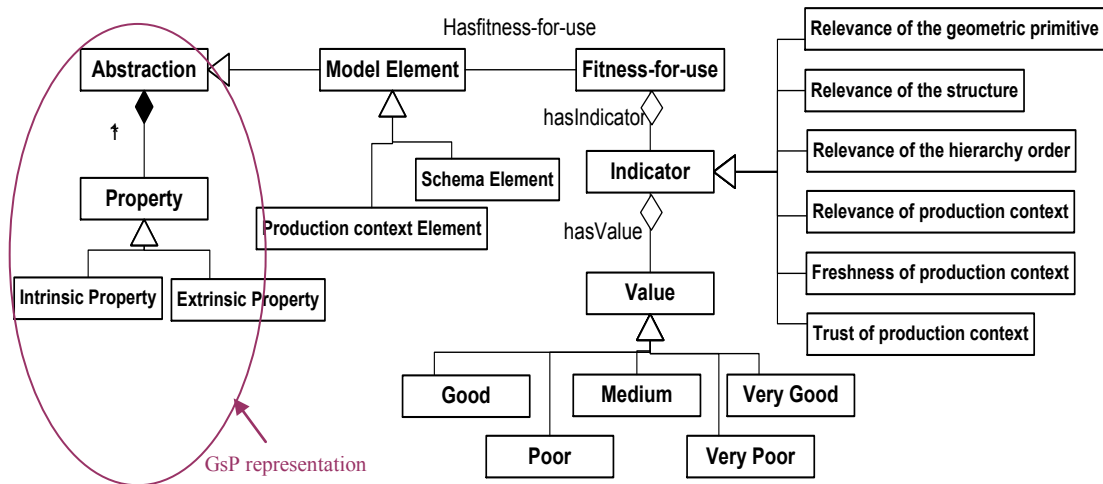


Figure 6.2: Embedding fitness-for-use representation within the representation of geospatial concept in GsP.

The next section presents the way the MQIGDC prototype for the interoperability between geospatial datacubes has been implemented.

6.3.2 MQIGDC implementation

6.3.2.1 Environment of implementation

The prototype MQIGDC was implemented using basically Java environment, XML format, and Perceptory tool. Java was used to construct the proposed conceptual framework (c.f., chapter 4) and to develop the method of evaluating the fitness-for-use of datacubes models (metadata and datacube schema). XML was used to represent message exchanged between software agents. Perceptory was used to develop geospatial repositories (i.e., ontologies). The choice of these technologies is motivated by:

1. The portability of the Java application.
2. The availability of Java libraries to process XML documents (namely the Java API for XML Processing (JAXP)) (JAXP, 2010) that includes the Xalan (The Apache Xalan Project, 2006) and the Xerces (The Appache Xerces Project, 2005) libraries for parsing and manipulating XML documents.
3. XML is widely recognized as a simple and flexible standard format that enables to exchange data between information systems. Data to be exchanged between geospatial datacubes is represented in XML documents.
4. The suitability of Perceptory tool to develop geospatial repositories agreeing to ISO19103 Geographic information–conceptual schema language (ISO/TC 211 2005a) and ISO19110 Geographic information–methodology for feature cataloguing (ISO/TC 211 2005b).

6.3.2.2 Interfaces of the MQIGDC prototype

The MQIGDC prototype implements six interfaces: *Agent manager interface*, *Datacube agent interface*, *Fitness-for-use interface*, *Quality evaluation interface*, *Warning & Suggestion interface* and *Decision-based communication interface*. The first two interfaces were taken from the GsP tool.

- *Agent manager interface* (Figure 6.3): it is used to instantiate software agents representing geospatial datacubes, and displays one agent's state upon user request. The instantiation of a datacube agent requires two elements: its identification and the name of the source ontology. Each agent can be instantiated by clicking the *New* button. The agent's state can be set active or inactive by clicking the *Start* or *Stop* buttons, respectively.



Figure 6.3: The agent manager window.

- *Datacube agent interface* (Figure 6.4): this interface aims to initiate the communication process between agents. In this interface contains two sections: the Console and the Communication Monitor. The Console section consists of three items. The first item is a dropdown menu, which presents the list of concepts that compose the agent's ontology. The second item is the *Send Query* button which initiates the communication process. The third item is a field in which the agent displays messages. The Communication Monitor section shows the processing of data exchange between agents.

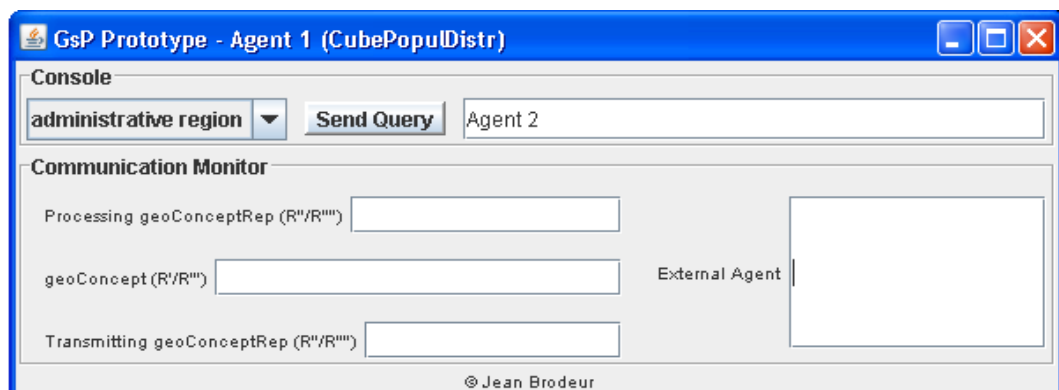


Figure 6.4: The agent window.

- *Fitness-for-use interface* (Figure 6.5): it is used to evaluate the fitness-for use of schema and metadata associated with each concept being exchanged between geospatial datacubes agents. This interface contains two main sections: the source context and the destination context. Both sections contain basically the same items, except the dropdown menu, which contains the list of the concepts matching the received concept, is contained only in the destination section. The first item is the source (or destination) context witch identifies the metadata ontology of the source (or destination). The stakeholder can enter an ontology name and click the *Refresh* button to display the ontology's elements within a dropdown menu. Once both ontologies are displayed, the stakeholder can choose an element (e.g., the method of aggregation and the measure function). Then, he/she can click the *Assess fitness-for-use* button in both sections to calculate the fitness-for-use. Also, this interface contains two other buttons: the *Context Similarity* button and *Fitness-for-use Comparison* button. The first button allows to measure the semantic similarity between the elements of context's concepts. The second button allows to compare the fitness-for-use of metadata and schema of both source's concept and destination's concept.

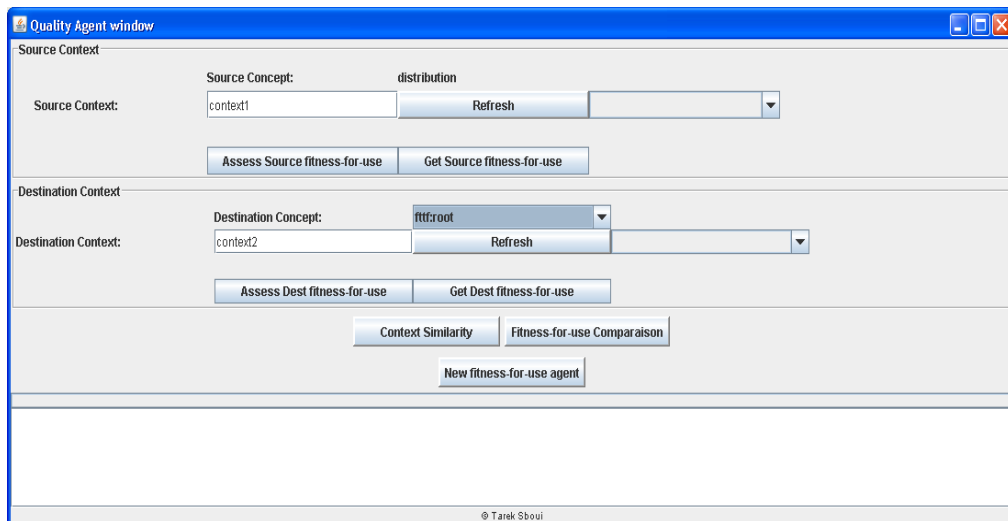


Figure 6.5: The quality (fitness-for-use) window.

- *Quality evaluation interface* (Figures 6.6 and 6.7): This interface is used to evaluate the quality of metadata and schema of the matched concepts. This

interface contains a quality indicator console and a two buttons: *Get indicator Value* and *Start Quality Assessment*. The quality indicator console contains six buttons that corresponds to the six indicators identified in our approach (c.f., chapter 5) as shown in figure 6.6.

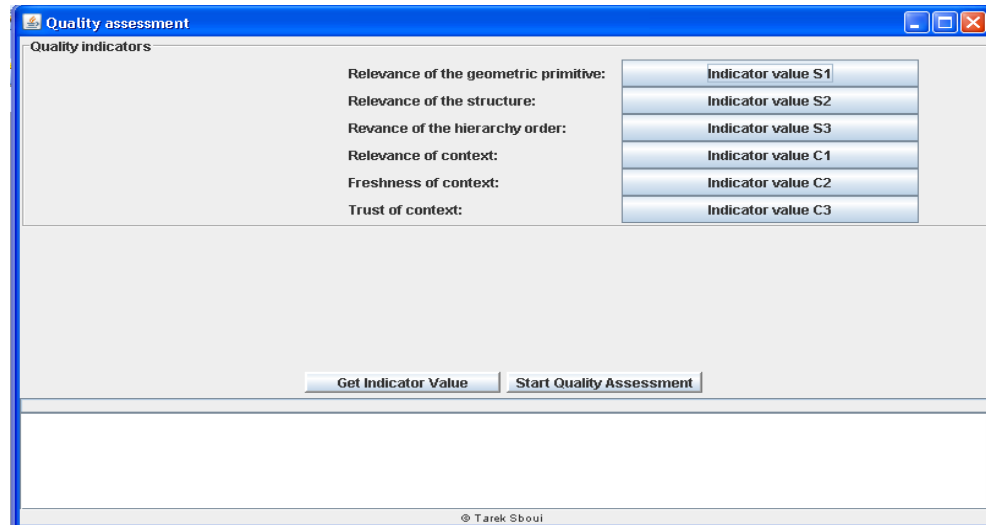


Figure 6.6: The fitness-for-use assessment window.

When the stakeholder clicks one of the six buttons, the *Calculation of indicator* window is displayed. This window allows to enter the parameters needed to calculate the indicator value according to the formula defined in chapter 5. For example, the window shown in figure 6.7 allows to enter the attribute of the formula (3) (c.f., chapter 5). Once the attributes have been entered, the prototype calculates the indicator according to the corresponding formula and stores the result in an XML document (using the *Calculate indicator* button).

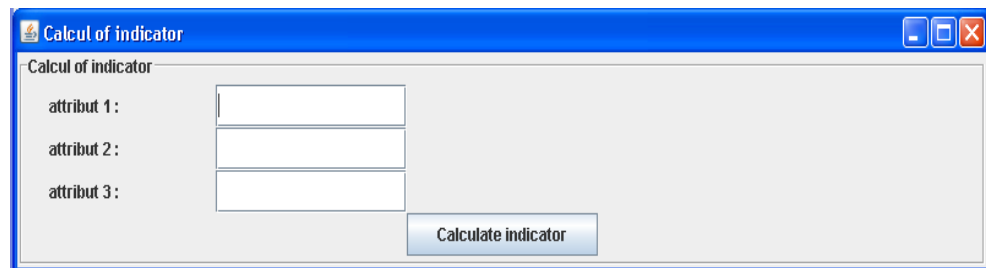


Figure 6.7: The indicator's value calculation interface.

The *Get indicator Value* button allows to obtain the quality value from the XML document and to display it. Then, the *Start Quality Assessment* button allows to evaluate the overall fitness-for-use and stores the result in an XML document.

- *Warning & Suggestion interface*: this interface consists of a warning symbol and a “suggestion” message such as the one shown in Figure 6.8. The warning symbol aims to make stakeholders aware of the potential risk of misinterpreting the exchanged concepts of geospatial datacubes. The message shows a suggestion driven from of the fitness-for-use assessment and formulated according to the algorithm defined in chapter 5. The message helps stakeholders to deal with the risks of data misinterpretation by proposing one or more actions that should be taken.

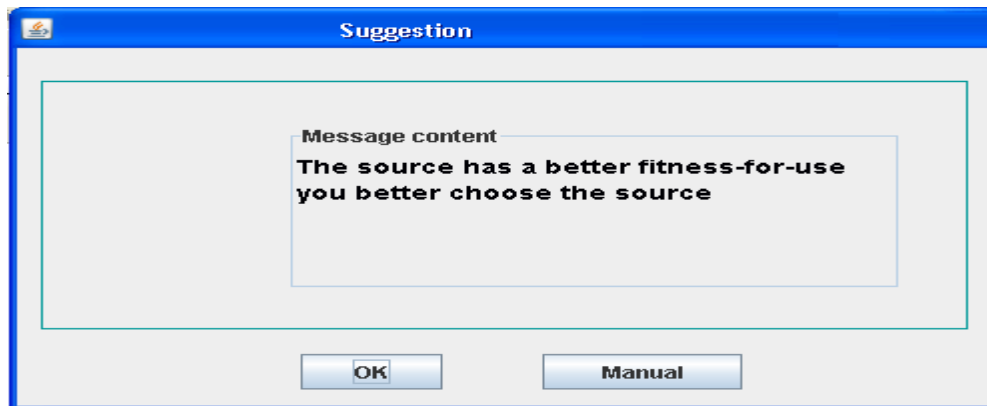


Figure 6.8: An example of a suggestion message.

- *Decision-based communication interface* (Figure 6.9): it contains three dropdown menus and three buttons. The two first dropdown menus contain the list of the destination agents, and the list of suggestions and warnings (e.g., “be careful, ignore the received message”). The third dropdown menu presents the list of possible responses of the destination agent to the source (e.g., “communication failed, the concept is not appropriate to my use”). The button *Suggest* allows the context agent to send a warning and a suggestion to the destination agent. The button *Send to more detailed level* allows the destination agent to send a warning and a suggestion to an agent representing a more

detailed level (of the same datacube). The button *Send final response* allows the destination agent to send his/her/its response to another datacube's agent.

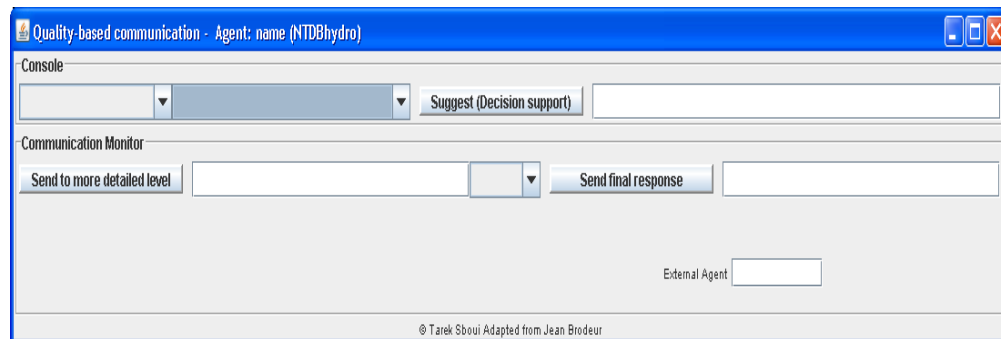


Figure 6.9 : Decision-based communication.

The communication of geospatial datacube agents starts by sending a message (e.g., a dimension's name "Time"). The message is encoded as a XML document. When another receives the message, MGsP tries to match the received concept to each concept of the destination ontology. Once the semantic similarity of both dimensions is measured, the context agent evaluates the fitness-for-use of their metadata and of their schema. For that, the MQSDCI prototype represents a set of indicators to the stakeholder. Each concept may have up to six quality indicators. In order to evaluate each indicator, the prototype uses the formulas defined in chapter 5. For each indicator, the stakeholder should introduce the value of each parameter needed to evaluate this indicator. For example, the stakeholder introduces 0.2, 0.5 and 1 as values of the following parameters defined in formula (3): N_{Elu} (the number of thematic, geospatial, or temporal elements of metadata, which are semantically related and required by end-users), N_{ReqElm} (the total number of required production context elements), and w (the level of importance of each type). Then, the overall quality associated to the current concept is evaluated using the formulas (6) and (7).

Once the overall quality is evaluated, and according to the algorithm defined in chapter 5, the context agent sends both the result (e.g., "the source has a better quality") and the suggestion (e.g., "you better choose the source") to the destination agent. Accordingly, the latter sends a feedback message to sender and a warning message to the agent representing the more detailed level of the destination datacube. Analyzing the warning message, the latter agent decides to proceed or not with the interoperability process.

In the next section, we present the experimentations conducted so far with the MQIGDC prototype.

6.3.3 Experimentation

In order to experiment the MQIGDC prototype, we continue with the example of geospatial datacube presented in Figure 3.8 (a and b). We remind that we use an extraction of two different geospatial datacubes developed for different analysis purposes. The first datacube is used to determine the distribution of the population in specific areas and periods (Bernier et al. 2009). The second datacube intends to analyze the risk of fire in Canadian forests according to a set of criteria (e.g., time and regions) (SOPFEU 2008). We also specify that the goal of interoperability, in our example, is to analyze the risk of forest fire on the Canadian population. For that, end-users would like to have a common model that allows them to navigate through both geospatial datacubes. We suppose that the requirements of end-users are represented as a geospatial datacube model, shown in Figure 5.2.

We should notice here that, we would have liked to experiment our approach with complete real cases, with users not involved in this research project. However, at the moment we began this experimentation (November 2008), this was not possible because at that moment, there has been only few geospatial datacubes in the industry (a large number of them were developed at the Centre for Research in Geomatics, Université Laval). The few existing geospatial datacubes, while have been developed based on good software engineering practices (e.g., using standard design language such as UML, and avoiding dependencies between datacube models and code), they have not been developed with reuse in mind (i.e., the aim of the interoperability). That is, each of these datacubes has been developed for a specific application, and has not been based on formal ontologies (e.g., using OWL). For instance, in our example, there was no ontology on which we can base our experimentation. This led us to define four ontologies for the geospatial datacubes of the example presented in section 3.4; two ontologies representing the content of the two geospatial datacubes, and two ontologies representing the production contexts related to these datacubes. These ontologies were defined, prior to the interaction process, based on geospatial datacube models, and then stored in four databases. Each of these databases

contains the concepts and properties of each ontology (intrinsic and extrinsic properties including multidimensional aspects such as method of aggregation and hyper-cellability). The defined ontologies are used to assign semantics to the elements of the two geospatial datacubes involved in the interoperability process.

In our experimentation, software agents were instantiated according to the geospatial datacubes of the Figure 3.8. For each datacube level (cube, measure, dimension, hierarchy, level), we define an agent. Moreover, we define a context agent to support the interoperability process.

We should notice that, in this experimentation, we suppose that the stakeholder will examine the warnings and accept the suggestions proposed by the MQIGDC prototype. Also, to evaluate the overall fitness-for-use and, for simplicity reasons, we set the value 1 for the variables a and b which indicate the importance of each quality indicator of schema and metadata respectively (c.f., equations (6) and (7)).

6.3.3.1 Phases of experimentation

The MQIGDC prototype involves the five layers of the proposed conceptual framework (cube, measure, dimension, hierarchy, and level) (c.f., chapter 3). Accordingly, the prototype's operation starts by initiating the communication between agents representing the cube layer of both geospatial datacubes going down to the level layer. At each layer, there is a decision to be made regarding the continuation or cessation of the interoperability. This decision is made collaboratively between agents of the same datacubes (representing different levels) and between the agents of the different datacubes.

We should notice that, since details about how the interoperability between geospatial datacubes is handled are provided in Annex D, in the following we focus on how the MQIGDC prototype provides support for such interoperability.

Cube layer:



The subjects of analysis of both datacubes are not semantically related. Thus, there is no conflict at the cube layer. The prototype displays the symbol  (see Figure 6.10).



Figure 6.10: Suggestion to continue to the next level.

Measure layer:

The agent representing the measure layer of *C1* communicates the name of the measure *Distribution* to the agent representing the measure layer of *C2*. The tool matches the measure *Distribution* with the measure *Fire-zone* by comparing their semantics based on the extended GsP (i.e., MGsP). The comparison results in MGsP_ffff (disjoint) predicate. This means that the measures have not been defined for the same subject of analysis. Accordingly, there is no semantic conflict at the measure layer. The stakeholder is invited to continue to the next layer (dimension layer). Then, the symbol  is shown.

Dimension layer:

The dimensions *Administrative region* of datacube *C1* and *Region* of datacube *C2* are similar (i.e., have MGsP_tfft (equal) predicate). In this case, to facilitate making a decision about the choice of one dimension over the other, the MQIGDC prototype evaluates the fitness-for-use of the schema and metadata of both dimensions with regard to the required dimension (i.e., *Administrative division*) (see Figure 6.11).

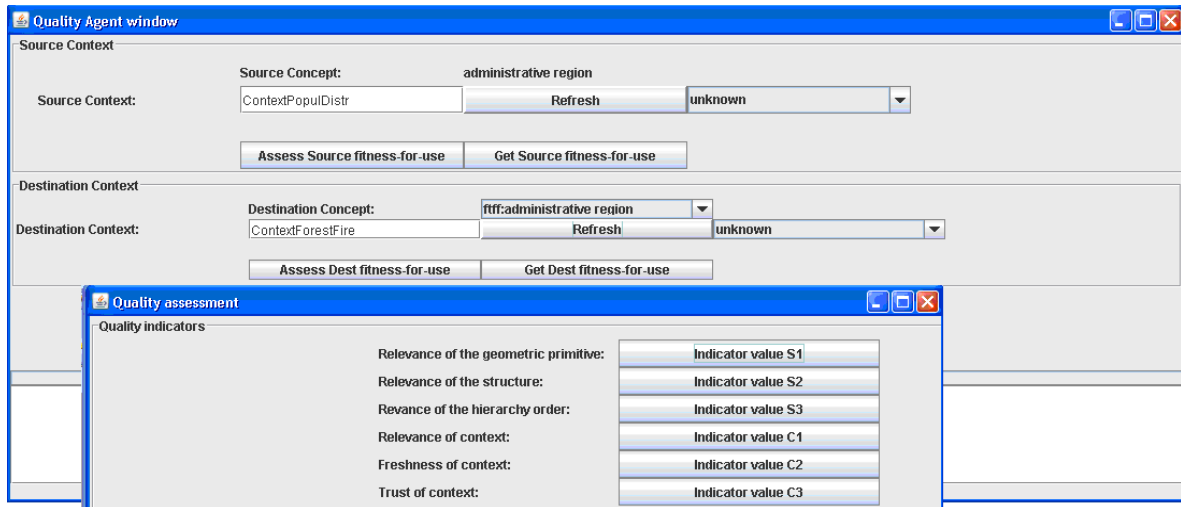


Figure 6.11: Evaluation of the fitness-for-use of the schema and metadata of the dimensions *Administrative region* and *Region*.

- Fitness-for-use of dimension's schema:

- *Relevance of the number of hierarchies (relevance of structure)*

According to the formula (1): $P_s(\text{Administrative region, Administrative division}) = P_s(\text{Region, Administrative division}) = 1/1 = 1$ (see Figure 6.12).

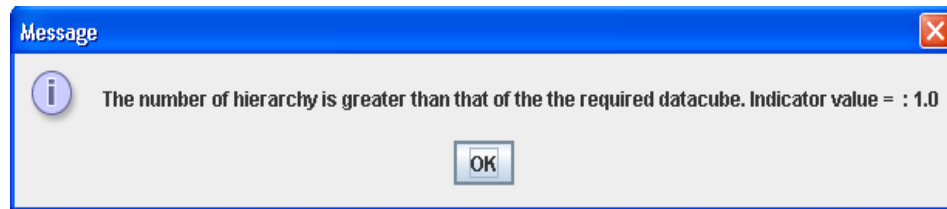


Figure 6.12: Value of the relevance of the number of hierarchy of *Administrative region*.

Then, the value of the overall fitness-for-use of schema Q_s is calculated according to formula (6): $Q_s(\text{Administrative region, Administrative division}) = Q_s(\text{Region, Administrative division}) = 1$ (see Figure 6.13).

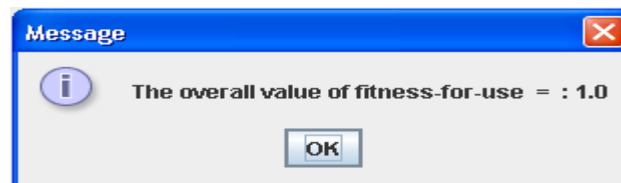



Figure 6.13: Value of the overall fitness-for-use of structure.

The prototype shows that the overall fitness-for-use of schema of both dimensions is very good. Thus, the risk of data misinterpretation is very low. Consequently, the

prototype displays the symbol  (see Figure 6.10) for both dimensions (with regard to their schemas).

- Fitness-for-use dimension's metadata:

➤ *Relevance of the metadata*

According to formula (1), the relevance of metadata is calculated as follows:

$$P_m (\text{Administrative region}, \text{Administrative division}) = 2/2 = 1 \text{ (Figure 6.14).}$$

$$\text{Similarly, } P_m (\text{Region}, \text{Administrative division}) = 1/2 = 0.5$$

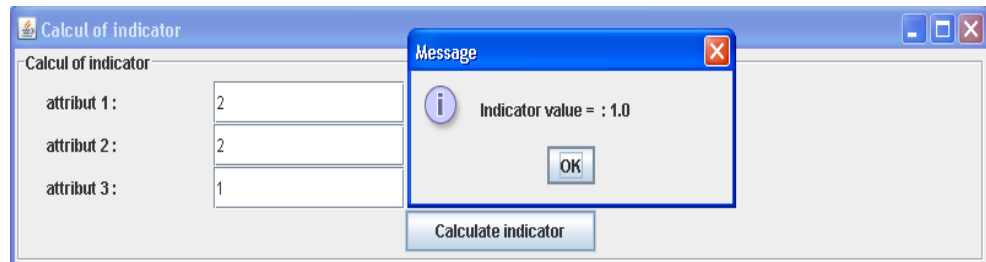


Figure 6.14: Value of the relevance of the number of hierarchy of *Administrative region*.

➤ *Freshness of metadata*

According to formula (4): $A_m (\text{Administrative region}, \text{Administrative division}) = 1 - (2005-2002/30) = 0.9$

Similarly, $A_m (\text{Region}, \text{Administrative division}) = 1 - (2005-1982/30) = 0.23$ (see Figure 6.15).

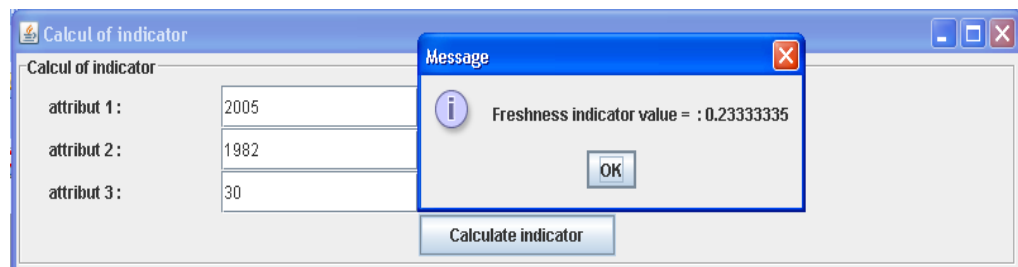


Figure 6.15: Value of the freshness of metadata of to the dimension *Region*.

According to formula (7), the overall fitness-for-use of metadata:

$Q_m(\text{Administrative region}, \text{Administrative division}) = (1 + 0.9) / 2 = 0.95$ (see Figure 6.16).

Similarly, $Q_m(\text{Region}, \text{Administrative division}) = (0.5 + 0.23) / 2 = 0.365$

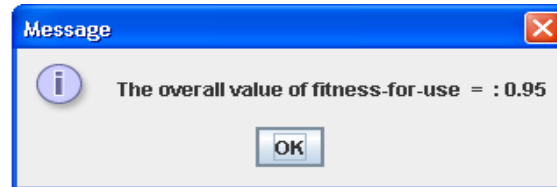


Figure 6.16: Value of the overall fitness-for-use of metadata related to the dimension *Administrative region*.

Consequently, the MQIGDC prototype displays the symbol **NOTICE** for the *Administrative region* dimension, and the symbol **WARNING** for the dimension *Region* (see Figure 6.17). The prototype (context agent) invites the stakeholder to be careful when considering the dimension *Region* in the interoperability.

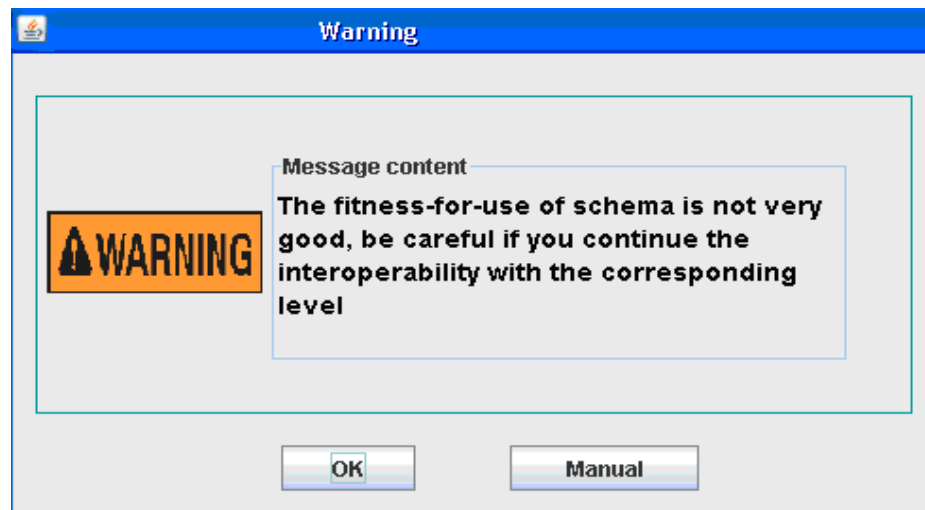


Figure 6.17: Making the stakeholder aware of the risk in considering the dimension *Region*.

Then, the context shows both the result of fitness-for-use and the suggestion (e.g., “be careful to consider the dimension *Region*”) to the source dimension agent (agent representing the dimension *Administrative region*). The latter sends a warning message to the more detailed agent (e.g., the agent representing the hierarchy; *H1*: *City*, *Province*, *Territory* and *Country*) (see Figure 6.18). In this case, the stakeholder is invited to be careful when considering the dimension *Region* in the process of interoperability.

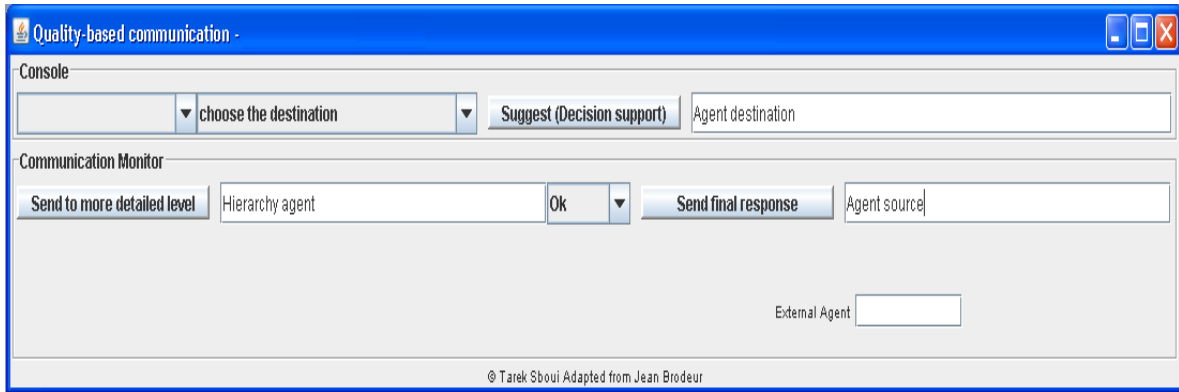


Figure 6.18: Context agent communicating with the dimension agents.

Hierarchy layer:

In the same way the tool calculates the fitness-for-use of the schema and metadata of the hierarchies (*H1: City, Province, Territory and Country*) of the dimension *Administrative region* and (*H2: City, Province, Territory and Country*) of the dimension *Region*.

- Fitness-for-use of hierarchy's schema:

- *Relevance number of levels (relevance of structure):*

According to formula (1):

$$P_S(H1, H3) = P_S(H2, H3) = 4/4 = 1$$


- *The order of levels*

According to expressions 2a and 2b (c.f., section 5.3.2.1):

$$O_S(H1, H3) = O_S(H2, H3) = (1+1+1+1) / 4 = 1$$

The overall fitness-for-use of the structure is calculated using formula (6):

$$Q_S(H1, H3) = Q_S(H2, H3) = 1$$

The overall fitness-for-use of structure of both hierarchies is very good. Accordingly, the risk of data misinterpretation is very low. Consequently, the MQIGDC prototype displays the symbol  (see Figure 6.10).

- Fitness-for-use of hierarchy's metadata:

➤ *Relevance of metadata*

We consider that the level of importance of spatial information is = 1, then, according to formula (3):

$$P_m(H1, H3) = 1/1 = 1$$

$$\text{Similarly, } P_m(H2, H3) = 0/1 = 0$$

The overall fitness-for-use of metadata is calculated according to formula (7):

$$Q_m(H1, H3) = 1/1 = 1$$

$$Q_m(H2, H3) = 0/1 = 0$$



Consequently, the MQIGDC prototype displays the symbol  for the hierarchy (*City, Province, Territory and Country*) of *C1* (see Figure 6.10). Also, the fitness-for-use of the metadata of the hierarchy (*City, Province, Territory, and Country*) of *C2* is very poor (0). Accordingly, the risk of data misinterpretation is high. Consequently, the MQIGDC prototype displays the symbol  for this hierarchy (see Figure 6.19). The prototype (context agent) invites the stakeholder not to consider the latter hierarchy since it may cause risks of data misinterpretation that may harm the interoperability. Thus, the prototype evaluates only the fitness-for-use of the levels of *H1*.



Figure 6.19 : Making the stakeholder aware of the risk of considering the hierarchy (*City, Province, Territory, and Country*) of *C2*.

Level layer:

In the same way the tool calculates the fitness-for-use of the structure and metadata of the level *City* of *CI*.

- Fitness-for-use of level's schema (Level *City* of *CI*):

- *Relevance of the number of attributes (relevance of structure):*

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

- *Relevance of the geometric primitive*

$$P_P(\text{City}, \text{City}) = 0.5$$

Then, according to formula (6), the MQIGDC prototype calculates the overall quality of the structure of the level *City* of *CI*:

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

This value is reasonably satisfactory, then MQIGDC prototype displays the symbol



(see Figure 6.17). Based on the proposed framework, the stakeholder can decide to solve the problems related to this level or to endure the potential consequences of these problems.


- Fitness-for-use level's metadata (Level *City* of *CI*):

- *Relevance of metadata*

$$P_m(\text{City}, \text{City}) = 2/2 = 1$$

According to formula (7), the quality of the metadata of the level *City* of *CI*:

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

The overall fitness-for-use of metadata of the level *City* of *CI* is very good. Accordingly, the risk of data misinterpretation is very low. Consequently, the MQIGDC prototype displays the symbol  (see Figure 6.10).

Similarly, the fitness-for-use of the levels *Province*, *Territory*, and *Country* of the datacube *C1* is evaluated.

6.3.3.2 Summary of the results

The table 6.1 shows the result of the evaluation of the fitness-for-use of schema and metadata related the remaining elements of both geospatial datacubes. Among these elements, the dimensions *Time* of the datacube *C1* and *Period* of the datacube *C2* are semantically related. Hence, they are handled the same way as the dimensions *Administrative region* of datacube *C1* and *Region* of datacube *C2*. As for the elements that do not cause a semantic conflict (i.e., they are not semantically related), they are integrated in the resulting common model if they fit requirements of end-users. It is the case for the dimensions *Age category* and *Forest stand*.

Element	Level	Datacube	Fitness-for-use of schema	Fitness-for-use of metadata
Time	Dimension	C1	1	1
Period	Dimension	C2	0.5	1
Year	Level	C1	1	1
Year	Level	C2	1	0.43
Month	Level	C1	1	1
Age category	Dimension	C1	1	1
Forest stand	Dimension	C2	1	1

Table 6.1: The fitness-for-use of datacubes elements.

If the stakeholder accepts the suggestions and warnings proposed by the MQIGDC prototype, then he/she will obtain a common model (e.g., the model presented in Figure D.1). Using such a common model would help end-users to navigate simultaneously through the geospatial datacubes *C1* and *C2*.

In this case study, the MQIGDC prototype helped the stakeholder to define a common model for both geospatial datacubes based on a set of fitness-for-use indicators of schema and metadata of geospatial datacubes. This experimentation result showed that:

- The context agent facilitates the communication between datacubes agents.
- Managing the risk of data misinterpretation based on the evaluation of the fitness-for-use of schema and metadata can fine-tune the semantic interoperability between geospatial datacubes.

Also, we should notice that the resulting model will support future communication between datacube agents. That is, in the future, the agents representing both datacubes will have an already-defined model which represents a common base for the agent to communicate without preoccupying themselves with the semantic conflicts that may occur.

The table 6.2 shows the result obtained if we test some concepts from Figure 3.8 with just the GsP tool proposed by Brodeur (first three columns of the table), and the one obtained using MQIGDC prototype (the entire table).

Element	Element	GsP	Evaluation result of fitness-for-use (using MQIGDC)	Suggestion proposed by MQIGDC
Administrative region (<i>C1</i>)	Region (<i>C2</i>)	GsP_tfft (equal)	<ul style="list-style-type: none"> - Qs (Administrative region) = 1 - Qm (Administrative region) = 0.95 - Qs (Region) = 1 - Qm (Region) = 0.365 	<p>For the dimension Administrative region: continue to the next level (hierarchy <i>H1</i>).</p> <p>For the dimension Region: you can continue to the next level (<i>H2</i>) but be careful the overall quality of metadata is poor.</p>
<i>H1</i> : City, Province, Territory, Country (<i>C1</i>)	<i>H2</i> : City, Province, Territory, Country (<i>C2</i>)	GsP_tfft (equal)	<ul style="list-style-type: none"> - Qs (<i>H1</i>) = 1 - Qm (<i>H1</i>) = 1 - Qs (<i>H2</i>) = 1 - Qm (<i>H2</i>) = 0 	<p>For the hierarchy <i>H1</i> : continue to the next level</p> <p>For the hierarchy <i>H2</i>: do not consider <i>H2</i> in the interoperability process (useless to continue to the next level 1).</p>
City (<i>C1</i>)	City (<i>C2</i>) (only in GsP tool)	GsP_tfft (equal)	<ul style="list-style-type: none"> - Qs (City) = 0.5 - Qm (City) = 1 	The stakeholder is informed of the quality. He/she can consider the element.

Table 6.2: Supporting the GsP tool to fine-tune the interoperability process.

The table 6.2 emphasizes the novelty of our approach with respect to the GsP approach. It shows how the MQIGDC prototype goes a step further to fine-tune the result of the matching process. In fact, even if the datacube concepts appear to be similar, the relevance of their schema and metadata for the intended use of the interoperability is different; one may be better than the other. The MQIGDC prototype aims, first, to make stakeholders aware of such differences that may cause a risk of data misinterpretation (e.g., if the stakeholder choose the concept that has the worse quality of schema and metadata), and second, to suggest some actions that allow stakeholders to manage the risks. This is especially beneficial in two situations:

- 1- The interoperability process between geospatial datacubes may require choosing between heterogeneous elements of these datacubes to create a new one. In such a case, it will be better to consider the element with a better quality value. Without such an evaluation, the stakeholder (software agent or a human stakeholder) may be uncertain about what concept he/she should choose. Such uncertainty may cause risks that may defeat the purpose for which the interoperability is carried out. The MQIGDC prototype evaluates the risks related to such uncertainty, and proposes relevant suggestions to the stakeholder (e.g., “it is better to choose the concept with a better quality of schema and metadata”).
- 2- When the schema and/or metadata of one of the matched concept is of very poor quality and hence not relevant to be considered in the interoperability outcome (e.g., an integrated model). The MQIGDC prototype makes the user aware of such irrelevance and suggests not to consider this concept.

The experimentation shows how the MQIGDC prototype fine-tunes the semantic interoperability of geospatial data by evaluating the risks of data misinterpretation, and by proposing some suggestions to support stakeholders in making appropriate decisions about such risks.

We should notice that in order to effectively manage the risks of data misinterpretation, people involved in the processes of data definition and design should define the information

needed to evaluate the fitness-for-use of schema and metadata. For example, in order to evaluate the freshness of metadata, data producer should provide metadata lifetime.

In addition, we believe that the results of this research will have impact on other research applications. For instance, the proposed indicators and method for managing the risks of data misinterpretation may be used in geospatial Web services. In this sense, we had interesting discussions with Dr Jean Brodeur about the application of the proposed approach to enhance the Web service we developed during this thesis's process for the North American Profile of ISO19115:2003 - Metadata (NAP – Metadata) registry. In fact, our approach would 1) help protect users of the NAP – Metadata registry from the risks of misinterpreting the registry's elements, and 2) enhance the retrieval of registry's elements by providing suggestions to choose those with a better quality value (especially when there is uncertainty about choosing among several semantically related registry's elements).

Furthermore, existing solutions of interoperability can extend their functionalities based on our prototype. For example, embedding the proposed risk indicators would help existing solutions to manage the risks of data misinterpretation, and hence to fine-tune the interoperability process in these solutions. These indicators can be used, for instance, to improve the efficiency of concept comparison routine in existing solutions of information retrieval. They can make end-users aware of potential risks of mis-correspondence between their search keywords and the retrieval results, or suggest to choose among several retrieval results (the result with a better quality value).

6.4. Conclusions

In this chapter, we presented the MQIGDC prototype that was developed to test and evaluate our proposed approach (i.e., the communication framework and the risk management approach). The prototype consists basically of 1) software agents representing spatial datacubes (i.e., datacube agents), 2) a context agent who supports datacube agents, 3) a method to define and to evaluate the risk of data misinterpretation by assessing the fitness-for-use of datacubes model elements, and 4) a method to support software agents or potential human stakeholders to make appropriate decisions about the defined risks. Agents communicate together by generating and transmitting geospatial concepts in XML. In order

to measure the semantic similarity between these concepts, we used the extended GsP tool (MGsP) on which our implementation is based.

The MQIGDC prototype was experimented using two geospatial datacubes modelled respectively in Figure 3.8 (a) and Figure 3.8 (b). The interoperability between the two datacubes aims to analyze the risk of forest fire on the Canadian population. The experimentation showed how the MQIGDC prototype supports interoperability of geospatial datacubes at different levels (cube, measure, dimension, hierarchy, and level) based on the mediation of a context agent, and on the evaluation of fitness-for-use of schema and metadata in these datacubes.

Although we consider that the MQIGDC prototype has been demonstrated by an experimental result of the interoperability between geospatial datacube (i.e., a common model that makes use of different geospatial datacubes: the distribution of the population (C1) and the forest fire (C2) datacubes), a number of issues still need to be addressed notably extracting metadata from natural language, and the ability to exchange a more complex message that contains a set of nominal and geometric data.

Finally, we believe that this research takes a step forward in fine-tuning the interoperability not only between geospatial datacubes, but also between other data sources. Also, some tasks which are manually performed can be automated. For example, instead of manually choosing the metadata element corresponding to a datacube element, it can be automatically discovered based on its relation with datacube elements in a combined ontology (i.e., ontology containing datacube elements associated and their metadata elements).

Chapter 7: Conclusions and discussion

7.1 Summary

Geospatial datacubes are useful for strategic decision-making. They allow decision makers to navigate aggregated and summarized data according to a set of dimensions with different hierarchies, and to gain insight into data by fast and interactive access to a variety of possible views of geospatial data. It might happen that one needs to reuse data stored in more than one existing geospatial datacube. Interoperability has been widely known as an efficient paradigm for enabling data reuse. Although the interoperability of information systems has attracted the attention of many researchers, no work on the interoperability of geospatial datacubes has been found. This thesis proposed an approach to support the semantic interoperability between geospatial datacubes. More specifically, the thesis discussed the need for interoperating geospatial datacubes, and proposed a categorization of semantic heterogeneity problem of geospatial datacubes (chapter 3). Then, it defined the characteristics of the semantic interoperability between geospatial datacubes, and presented a conceptual framework which is based on software agents to support this interoperability (chapter 4). Then, it explained the risks of data misinterpretation related to this interoperability and proposes an approach to manage them (chapters 4 and 5). This approach is implemented and tested to prove its contribution (chapter 6).

7.2 Contributions

In chapters 3-6, we presented the contributions of our research work. These contributions to the theoretical corpus specific to geospatial datacubes resulted into nine peer-reviewed papers published in international conferences and journals (Sboui et al. 2007, Sboui et al. 2008, Sboui et al. 2009a, Sboui et al. 2009b, Sboui et al. 2009c, Sboui et al. 2009d, Sboui et al. 2010a, Sboui et al. 2010b, Sboui et al. 2010c). The main contributions of this thesis are:

- 1 – *Determining the need for the interoperability between geospatial datacubes.* Determining the need for the interoperability between geospatial datacubes is an important

step to define such interoperability. Thus, we first discussed situations where there is a need for interoperating geospatial datacubes. We grouped these situations into three categories:

1. A simultaneous and rapid navigation through different datacubes, where users need to access and navigate simultaneously through heterogeneous geospatial datacubes. In such situation, interoperability would enable, for example, to rapidly navigate through data stored in the geospatial datacubes.
2. A rapid insertion of data in a datacube, where users need to rapidly insert new data in a geospatial datacube from other datacubes. Such insertion may require exchanging data between geospatial datacubes, i.e., the aim of interoperability.
3. An interactive and rapid analysis of change. In order to analyze change (e.g., forest stand dynamics), we may need to compare data describing these phenomena at different epochs. Interoperating geospatial datacubes would enable to interactively compare data and analyze change.

Our first contribution was published in (Sbouï et al. 2007).

2 – Defining the interoperability between geospatial datacubes. While the interoperability between transactional databases and the interoperability between datacubes have a common *raison d'être* which is data reuse, the latter has particular characteristics due to the specific structure and elements of datacubes. However, our literature review showed that no definition of the interoperability between geospatial datacubes has been proposed. Defining the characteristics of such interoperability between geospatial datacubes constitutes another contribution of this thesis. We defined the interoperability between two geospatial datacubes C1 and C2 as the ability of C1 to request a service in a manner that can be understood by C2, and the ability of C2 to respond to that request in a manner that can be understood by C1 and adapted to its context. A service may involve one or more of the following categories of actions:

1. Comparing an element of a geospatial datacube (i.e., measure, dimension, hierarchy, or level) against an element of another geospatial datacube.

2. Updating an element of a geospatial datacube based on the content of other datacubes involved in the interoperability process. This may include modifying one or more semantic aspects of the datacube element (e.g., modifying the definition of a measure, changing the context of a dimension). It may also include updating geospatial members, updating member properties, or updating the values of some measures.
3. Integrating datacubes involved in the interoperability process. In this case, the interoperability between datacubes would lead to the creation of one or more federated geospatial datacubes. This refers basically to the creation of a common conceptual model from the models of the datacubes involved in the interoperability process. Such a common model would allow to virtually navigate through existing datacubes as if they were only one. This category may involve integrating measures or integrating dimensions of geospatial datacubes.

We also proposed a categorization of semantic heterogeneity problems that may occur during such interoperability. The defined categories are: Cube-to-Cube heterogeneity, Measure-to-Measure heterogeneity, Dimension-to-Dimension heterogeneity, Hierarchy-to-Hierarchy heterogeneity, and Level-to-Level heterogeneity. In each category, we distinguished between schema heterogeneity and context heterogeneity.

Our second contribution was published in (Sboui et al. 2007), (Sboui et al. 2008) and (Sboui et al. 2010b).

3 – Developing a specific framework for the interoperability between geospatial datacubes.

Our literature review revealed that while existing approaches of semantic interoperability between transactional databases can be used to a certain extent to support the interoperability between geospatial datacubes, the efficiency of such interoperability can be improved by developing a framework specific to geospatial datacubes. Accordingly, another contribution of this thesis is to propose a framework to support the interoperability between geospatial datacubes. The framework is based on two kinds of software agents: *datacube agents* and a *context agent*. Datacube agents represent geospatial datacubes to be involved in the interoperability process, and are engaged in a process of human-like

communication. They are defined according to five layers that correspond to the five conceptual elements of datacube: cubes, measure, dimension, hierarchy, and level. Context agent makes datacube agents aware of potential misinterpretation related to the inappropriateness of context information or of schema.

The proposed conceptual framework defines two types of communication: 1) horizontal communication between agents representing the same layer belonging to different geospatial datacubes, and 2) a vertical communication between agents of the same datacube at different layers. The hierarchical structure of the proposed framework corresponds to the proposed categorization of semantic heterogeneity. Hence, it allows agents to resolve the semantic heterogeneity occurred in each category at once; starting from the highest and going down to the lowest layer of each datacube. This permits to reduce the complexity of the overall semantic problems.

Our third contribution was published in (Sboui et al. 2007).

4 – *Extending the geosemantic proximity*. In order to compare and match data communicated between agents, we based our framework on the geosemantic proximity approach (GsP). This approach consists of evaluating the semantic similarity between geospatial concepts (i.e., similarities between their intrinsic and extrinsic properties). Moreover, in order to deal with particular semantic heterogeneities of geospatial datacubes, we proposed an extension of the GsP approach to include the comparisons of basic multidimensional concepts such as semantics of aggregation and semantics of hypercellability. The objective of this extension (termed MGsP) is to give an agent the possibility to dig into more detail about the semantic heterogeneity of important notions of the multidimensional paradigm.

Our fourth contribution was published in (Sboui et al. 2010a).

5 – *Managing the risks of data misinterpretation associated with the interoperability between geospatial datacubes*. The MGsP relies on ontologies to measure the semantic similarity between geospatial concepts. However, automatic ontology-based matching still suffers from problems related to the quality of matching results. Moreover, data description

may have different interpretations which are more or less relevant to the intended meaning (semantic uncertainty). These result in risks of data misinterpretation. Existing approaches of interoperability tried to deal with such a risk in a non-systematic manner (i.e., not based on predefined and ordered criteria). Consequently, semantic interoperability still remains vulnerable to the risks of data misinterpretation. Such risks may affect strategic decision-making process when interoperating geospatial data cubes. Accordingly we proposed a systematic approach to deal with the risks of data misinterpretation. The approach is based on evaluating the fitness-for-use of the production context and of the schema of geospatial data cubes. In fact, a production context and a schema which are appropriate for an intended use may be of poor fitness-for-use for data reuse (i.e., the use for which the interoperability is carried out). Evaluating the fitness-for-use of schema and production context gives a clue about the risks of misinterpreting geospatial data cubes content and can help protect end-users from such risks. Indeed, while a good fitness-for-use indicates it is less likely to faulty interpreting data or being uncertain about its interpretation, a poor quality indicates a higher risk.

Managing the risks of data misinterpretation represents a novelty in the field of interoperability. Compared to existing approaches, such as the geosemantic proximity (GsP), our approach goes a step further to fine-tune the interoperability process by evaluating the relevance of schema and production context with respect to the application for which the interoperability is carried out.

We also introduced a set of indicators to evaluate the fitness-for-use of production context information and data cube schema. These indicators play an important role in managing the risks of data misinterpretation. They enable the identification and the evaluation of these risks, and hence, make users aware of their severity. These indicators are: relevance of the geometric primitive, relevance of the structure, relevance of the hierarchy order, relevance of production context, freshness of production context and trust of production context. 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. Based on this value, a

qualitative value (i.e., “good”, “medium” or “poor”) may be assigned to the fitness-for-use. The definition and the evaluation of these indicators take into account users requirements.

We also proposed a general framework that can be used by stakeholders to make appropriate decisions about the risks that may occur during the interoperability process of geospatial datacubes. The framework consists of five phases of analyzing and responding to these risks. These phases correspond to the five layers of our conceptual framework of the interoperability between geospatial datacubes, from the more general layer to the most detailed: cube, measures, dimensions, hierarchies, and levels. As such, stakeholders can make appropriate decisions at each phase of the framework by taking into account their observations at early phases. At each phase, the evaluation result of the proposed indicators is represented qualitatively using warning symbols in order to make stakeholders aware of the risks of data misinterpretation, and to help them making decisions in an intuitive way.

We should remember that the set of the proposed indicators and the proposed framework do not aim at being exhaustive or precise but rather at helping the stakeholder to make appropriate decisions to enhance the interoperability between geospatial datacubes.

Our fifth contribution was published in (Sboui et al. 2009a), (Sboui et al. 2009c), (Sboui et al. 2009d) and (Sboui et al. 2010b).

In addition to the main contributions, there are two other secondary contributions of this thesis. The first contribution is a semantic model, called *SemEL*. This model enables to explicitly represent data meaning, and to help human stakeholders to intuitively interpret data according to ontology, spatio-temporal characteristics, and context. The *SemEL* model was the main subject of the paper titled “A Conceptual Framework to Support Semantic Interoperability of Geospatial Datacubes” (Sboui et al. 2007). An extraction of this article can be found in Annex B of this thesis.

The second contribution was to propose a way to implement the proposed indicators of the fitness-for-use of context using existing Semantic Web technologies such as OWL. This allows to define a number of rules to facilitate automatic context reasoning. More details about this work can be found in the paper titled “Modeling the External Quality of Context

to Fine-tune Context Reasoning in Geospatial Interoperability” (Sboui et al. 2009b). A modified version of this paper is available in Annex C of this thesis.

7.3 Discussion

This thesis deals with a subject that has never been tackled, i.e., interoperability between geospatial datacubes. It categorizes the problems that may occur when interoperating geospatial datacubes and proposes a new approach to deal with such problems. In order to develop the prototype of the proposed approach, we used and extended the GsP tool developed by Brodeur (2004). We consider our approach as a complementary to existing approaches as it allows them to fine-tune the semantic interoperability between databases by managing the risks of data misinterpretation in a systematic way.

We believe that the hypotheses made in this research work, which are 1) the possibility to provide an approach that supports the interoperability between geospatial datacubes and 2) the possibility to reduce the risks of data misinterpretation during such interoperability, have been verified. The hypotheses were validated (1) by identifying a conceptual framework of geospatial datacubes interoperability, (2) by identifying a systematic approach to manage the risk of data misinterpretation during the interoperability process, (3) by proposing a method to evaluate the fitness-for-use of schema and production context, and (4) by developing a prototype to experiment and test the proposed approach.

The thesis supports the vision of the interoperability of geospatial data based on the process of human-like communication and the cognitive model of humans. Also, it recognizes the risks related to such interoperability (i.e., risks of misinterpretation) which may occur during the communication process since the exchanged data can be employed in a different context than the one it was intended to be used in. In addition, the thesis proposes to integrate a mediator agent (i.e., context agent) that embeds relevant information about context with the exchanged data. That is, information about the fitness-for-use of context information and of datacube schema. Such information allows agents communicating together (or a human stakeholder) to make appropriate decisions to respond to the risk of data misinterpretation.

We noticed that risk management, context, and context quality are relevant elements that influence decisional geospatial data exchange. We showed that when these elements are combined with other elements such as human-like communication process, ontologies and semantic similarity, they can enhance the interoperability between geospatial data cubes.

We believe that the results of this research enhance decision support systems. In fact, users may efficiently reuse simultaneously more than one geospatial data cube which may be required for their strategic decision making process (i.e., simultaneously and rapidly navigation through different data cubes to manage forest fire located at the border area of two adjacent countries). This is done by making geospatial data cubes interact together and by facilitating a possible human intervention.

Moreover, we believe that the results of this research enhance the efficiency of the interaction between geospatial data providers and end-users. In fact, receiving data from a provider, end-users may faulty interpret this data or be uncertain about its intended meaning causing risks that may affect data use. Managing such risks may be required in many situations in which a data provider supplies data to end-users. An example of situation occurs when data provider is no more available to give more information (or respond) to end-users requests. Even if the provider is available, responding to end-users may be costly, in terms of monetary and time loss and/or human effort. In such situations, our proposed approach makes end-users aware of the risks related to the interpretation of received data, and helps them to deal with these risks.

Furthermore, we believe that the proposed approach can be used to support other works related to the interoperability of information systems in general. For instance, based on our approach, existing tools can extend their functionality to manage the risks of data misinterpretation, and hence to fine-tune the interoperability process. For example, the efficiency of concept comparison routine in information retrieval applications can be improved by managing related risks (e.g., risk of mis-correspondence between concepts). Also, our approach can be used to improve Semantic Web services by identifying and managing the risks related to data misinterpretation during the interaction between Web applications.

However, it is important to note that the proposed approach is not perfect; i.e., it does not resolve all the problems related to the risks of data misinterpretation. In fact, the approach is limited to the identification and evaluation of the risks based on the fitness-for-use of schema and production context. However, the risks of data misinterpretation can be caused by other elements such as the vagueness of geospatial data. While it is out of the scope of our research work, the study of such elements would be a significant addition to the proposed approach.

7.4 Final conclusions

Geospatial datacubes have proven their efficiency in strategic decision-making. It might happen that one needs to reuse the content of more than one existing geospatial datacube. Interoperability has been widely known as an efficient paradigm for enabling data reuse.

In this thesis, we developed an approach to support the semantic interoperability between geospatial datacubes. The approach proposed a categorization of semantic heterogeneities in geospatial datacubes. Then, it presented a conceptual framework which is based on software agents: datacube agents and a context agent. Datacube agents represent geospatial datacubes to be involved in the interoperability process. Context agent makes datacube agents aware of potential inappropriateness of context information. In addition, and in order to identify and evaluate the risks of data misinterpretation, the approach proposed a set of indicators to evaluate the fitness-for-use of schema and production context of geospatial datacubes. The approach also proposed a general framework that can be used by stakeholders to make appropriate decisions about the risks that may occur during the interoperability process of geospatial datacubes. Therefore, our works represents a meaningful step toward fine-tuning the interoperability of geospatial datacubes in particular and of information systems in general.

Besides, from the results of this thesis we can draw the following conclusions:

- In many situations, users may need to reuse more than one geospatial datacube (e.g., to rapidly navigate through different geospatial datacubes developed in adjacent jurisdictions during a natural disaster, to add measures about the

numbers of employees to a datacube *human resources* from a datacube *employee assignment*, to add the area of lakes and the area of roads in a geospatial datacube about the construction of winter bridges, from two other geospatial datacubes; *traffics* and *lakes*, and to rapidly analyze changes in the wood volume following a natural disaster by comparing information about forest stand stored in different geospatial datacube). Interoperating such datacubes enhance the efficiency and effectiveness of their reuse. Interoperating geospatial datacubes may be even required in some situations, such as forest fire in adjacent countries, where appropriate data has to be quickly obtained, and taking appropriate decisions have to be quickly and reliably taken.

- The proposed conceptual framework of interoperability (c.f., chapter 3) represents a model of how the interoperability between information systems can be performed.
- The proposed categorization of semantic conflicts was defined according to the structure of geospatial datacubes. Such categorization helps to deal with the semantic conflicts related to the interoperability between geospatial datacubes.
- The proposed approach of managing the risk related to the interoperability between geospatial datacubes constitutes a systematic management of the risks of data misinterpretation (i.e., identifying, evaluating, and responding to such risks).
- Defining a set of indicators of the fitness-for-use of production context information and of schema constitutes an efficient way to identify and evaluate the risks of data misinterpretation during the interoperability process.
- Qualitatively evaluating the fitness-for-use of context information based on a quantitative method constitutes an efficient and intuitive way to be aware of the risks of data misinterpretation and to help stakeholders to making decisions.

7.5 Research perspectives

The research work presented in this thesis represents another step towards the realization of a more efficient interoperability between information systems. It opens new research perspectives in the realm of databases interoperability, in general, and in the geospatial datacubes interoperability, in particular. Some of these research perspectives that deserve further explorations are presented below.

- *Relationships with other works.* The results of this thesis can be extended with the results of Mehrdad Salehi's and Lotfi Bejaoui's thesis (Salehi 2009, Bejaoui 2009). Mehrdad Salehi defined a formal model for geospatial datacubes and based on this model identified the integrity constraints of geospatial datacubes. These integrity constraints are important elements that describe the semantics of geospatial datacubes content. In addition, the proposed formal model enables to systematically and precisely document the integrity constraints in geospatial datacube applications at the design stage, hence, allowing reducing the vagueness of these constraints which is considered one of the causes of the risks of data misinterpretation. Consequently, these risks will be reduced. Lotfi Bejaoui proposed an approach to represent geospatial objects with partially or totally vague shapes, and to integrate them based on a set of topological relationships such as "weakly" "overlap", "strongly disjoint", and "completely equal". The approach allows users to have knowledge about the uncertainty of vague geospatial objects and of their relationships, and hence to make appropriate decisions about data interpretation and use. Consequently, the risk of misinterpreting the available data can be considerably reduced.
- *Considering other situations where the interoperability is needed.* Although we consider the prototype and the experimentation to be successful, a number of improvements may be made. One important improvement will be to consider other possible situations where the interoperability of geospatial datacube is needed (e.g., comparing geospatial datacubes content and inserting data in a geospatial datacubes from another one) rather than considering only one situation as it is the case in our experimentation (i.e., comparing geospatial datacubes

content to create a common model that allows a simultaneous and rapid navigation through different datacubes). Such an improvement must provide the choice of a specific situation for real end-users, i.e., end-users must be able to adapt the functionality of the future tool to their needs (e.g., comparing data stored in different geospatial datacubes). Such tool will greatly enhance the efficiency of geospatial datacubes interoperability.

- *Testing the approach in several other applications.* In order to test our approach, we used an extraction from the content of two different geospatial datacubes developed for different analysis purposes (distribution of the population and control of the forest fire extent). Moreover, we made some assumptions in order to avoid issues which are out of scope of this thesis (e.g., supposing that the user requirements are represented in the form of a geospatial datacube model, supposing that each datacube has its a predefined ontology, and setting the value 1 for the importance of each quality indicator, although they may have different values between 0 and 1). Such assumptions may not materialize in reality. A future step will be to test the approach in several real applications using different datasets without making such assumptions. Such tests will allow to further demonstrate the potential of our approach in supporting the interoperability of geospatial datacubes.
- *Evaluation of spatio-temporal aspects.* The proposed set of indicators proposed in this thesis take into account geospatial context (e.g., coverage region) and temporal context (year of data definition). However, this set does not take into account spatio-temporal context. Such context contains concepts that refer to both space and time. Examples of such concepts are the spatiotemporal primitive “moving-point” or the spatiotemporal concept “speed”. It would be beneficial that the proposed set of indicators in this thesis be extended to take into account spatio-temporal context.
- *Semantic of geometries.* In this thesis, the semantics of the geometry was restricted to geometric primitives. However, other geometric aspects may affect the semantics of data in geospatial datacubes, such as the form and the

orientation. Considering such aspects would enhance the reasoning about the semantics of geospatial datacubes content and facilitate their interpretation.

- *A specific tool for SemEL.* While *SemEL* can be instantiated using OLAP or SOLAP tools such as JMap SOLAP tool, developing a specific tool to edit and manage *SemEL* concept is an interesting research topic (e.g., a tool that has functionalities similar to ontology editing tools such as Protégé 2000). Such a tool would make human intervention more efficient and more effective. For example, such a tool can provide to users a list of functions to aggregate the semantics of geospatial datacubes content (i.e., ontology, context, and spatio-temporal elements).
- *Using the indicators to evaluate the risks of public use of geospatial data.* The proposed set of indicators and the method of evaluating them to identify and evaluate the risks of data misinterpretation can be used by the ongoing research project conducted by the Ph.D. student Joel Grira at Laval University which aims to evaluate the risks of public use of geospatial data (Computer-Assisted Risk Evaluation For Usage Limitations). This project can adapt the introduced set of indicators and reuse the proposed method to evaluate them.
- *Coupling quantitative and qualitative approaches.* Qualitative indicators are generally simple to be used and provide a coarse evaluation of risks of data misinterpretation. These indicators provide an intuitive way to identify and assess such risks. However, quantitative indicators provide a fine computation of risks. For example, they can indicate how much a context is more appropriate than another. In this thesis, we considered a qualitative result of evaluating indicators (based on quantitative functions). Combining both qualitative and quantitative results would be beneficial to users. Both results of evaluating indicators (quantitative and qualitative) can be represented in a framework where the qualitative evaluations are placed at its high level and quantitative ones in the bottom level. As such, it is possible, for example, to infer qualitative evaluations (e.g., “better”, “worse”) based on quantitative evaluations in a lower level. The user may have the choice to use qualitative evaluations or to drill-down in the

detail and use quantitative ones. Thus, such framework would provide the facility of qualitative evaluations and the precision of quantitative ones.

- *Considering the propagation of risks within the chain of stakeholders.* In the context of interoperability, the probability and the potential harm of data misinterpretation or data uncertainty constitute a risk which may occur each time data are interpreted. That is, the risk impacts may affect not only software agents communicating together, but may also propagate along the chain of stakeholders in the interoperability process (e.g., software agents, human/software mediator, user, and lawyer). Risk propagation becomes more complicated with the increase of the number of stakeholders in the process of the interoperability and may raise suspicion about the convenience of re-using data (the main aim of interoperability). Figure 7.1 shows a possible scenario of risk propagation. Agent 1 and Agent 2 are two software agents involved in a current interoperability process. Data of Agent 1 have been directly acquired from a designer. Data of Agent 2 have been integrated from other sources. Moreover, the interoperability chain involves an agent mediator who may have recourse to a domain expert to oversee data interpretation. Both agents and mediators may contact a lawyer in order to get advice on problems that can be legal matters. The study of the risk propagation was out of the scope of our research work. Studying such propagation will be important to manage the risks of data misinterpretation.

- *Testing the approach in other domains.* We suggest testing this approach in other domains such as urbanism, forestry, pollution, climatic changes, etc. The same prototype may be used to experiment our approach in these applications.
- *Enriching the proposed indicators.* While, the proposed indicators play an important role in responding to the risks of data misinterpretation, we remind that they do not aim at completely eliminating such risks. These indicators represent a step forward in identifying and evaluating the risks. While it is very difficult (even impossible) to define all possible indicators, the defined set in this work can be enriched by other indicators. Such addition will enhance the efficiency of risk identification and evaluation.

- *Testing the approach in other domains.* We suggest testing this approach in other domains such as urbanism, forestry, pollution, climatic changes, etc. The same prototype may be used to experiment our approach in these applications.
- *Enriching the proposed indicators.* While, the proposed indicators play an important role in responding to the risks of data misinterpretation, we remind that they do not aim at completely eliminating such risks. These indicators represent a step forward in identifying and evaluating the risks. While it is very difficult (even impossible) to define all possible indicators, the defined set in this work can be enriched by other indicators. Such addition will enhance the efficiency of risk identification and evaluation.

Bibliography

- Aalders, H.J.G.L., *The Registration of Quality in a GIS*, Spatial Data Quality, Shi, W., Fisher, P. and Goodchild, M.F. (Eds.), Taylor & Francis, 2002. p 186-199.
- Abbott R., *Subjectivity as a Concern for Information Science: a Popperian Perspective*, Journal of Information Science, 30(1), 2004. p 95-106
- Abelló, A., Samos, J. and Saltor, F., *YAM²: A Multidimensional Conceptual Model Extending UML*, Information Systems, 31(6), 2006. p 541-567.
- Afila, D. and Smith, NJ., *Risk management and value management in project appraisal*, Management, Procurement and Law, 2007 - Thomas Telford.
- Agumya, A. and Hunter, G. J., *Determining fitness for use of geographic information*, ITC Journal 2, 1997. p 109-113.
- Agumya, A. and Hunter, G.J., *Responding to the consequences of uncertainty in geographical data*, International Journal of Geographical Information Science, 16(5), 2002. p 405-417.
- Akoka, J., Berti-Equille, L., Boucelma, O., Bouzghoub, M., Comyn-Wattiau, I., Cosquer, M., Goasdoué, V., Kedad, Z., Peralta, V., Nugier, S., and Si-Said, S., *A framework for quality evaluation in data integration systems*, Proceedings of ICEIS'07, 2007.
- Albrecht, J., *Geospatial information standards a comparative study of approaches in the standardisation of geospatial information*, Computers & Geosciences (25), 1999. p 9-24.
- Allen, J. F., *An interval-based representation of temporal knowledge*, In Proceedings of international joint conference on artificial intelligence, 1981. p 221-226.
- Allen, J. F., *Maintaining knowledge about temporal intervals*, Communication of the ACM, 26, 1983. p 832-843.
- ANSI Z535.4, American national standard for product safety signs and labels, 1991.
- Atkinson, C. and Kühne, T., *Model Driven Development: a Metamodeling Foundation*, IEEE Transactions on Software Engineering (20), 2003. p 36-41.
- Azouzi, M., *Suivi de la qualité des données spatiales au cours de leur acquisition et de leurs traitements*, Thèse de doctorat en sciences techniques EPFL, Ecole technique polytechnique fédéral de Lausanne, 2000.
- Badard, T. and Braun, A., *OXYGENE: A Platform for the Development of Interoperable Geographic Applications and Web Services*. In proceedings of the 15th International Workshop on Database and Expert Systems Applications (DEXA'04), IEEE Press, Zaragoza, Spain, August 30 - September 2004. p 888-892.
- Bailin, S.C. and Truszkowski, W., *Ontology negotiation between intelligent information agents*. Knowledge Engineering Review, 17 (1), 2002. p 7-19.
- Ballou, D.P. and Tayi, G.K., *Enhancing Data Quality in Data Warehouse Environment*, Communication of the ACM 42(1), 1999. p 73-78.
- Batini, C., Ceri, S. and Navathe, S.B., *Conceptual database design: An Entity Relationship approach*, Benjamin Cummings, 1992.
- Beaulieu, V., *Étude de la visualisation géographique dans un environnement d'exploration interactive de données géodécisionnelles adaptation et améliorations*, Master's thesis, Laval University, 2009.
- Beard, K., *Use error: the neglected error component*, In Proceedings of Autocarto 9, Baltimore, MD, USA, 1989. p 808-817.

- Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. and Stein, L.A., *OWL Web Ontology Language Reference*, version 1.0. W3C, Technical report, W3C, <http://w3.org/TR/owl-ref/>, 2004.
- Bédard, Y., *A Study of Data using a Communication based Conceptual Framework of Land Information Systems*, *Le Géomètre Canadien*, 40 (4), 1986. p 449-460.
- Bédard, Y., *Uncertainties in Land Information Systems Databases*, *AUTO-CARTO* 8, 1988. p 175-184.
- Bédard, Y. and Han, J., *Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery*, *Geographic Data Mining and Knowledge Discovery* (2nd edition), Miller, H.J. and Han, J. (Eds.), Taylor & Francis, 2008.
- Bédard, Y. and Larrivée, S., *Spatial Database Modeling with Pictogrammic Languages*, *Encyclopedia of GIS*, Shekhar, S. and Xiong, H. (Eds.), Springer-Verlag, 2008. p 716-725.
- Bédard, Y., Rivest, S., Proulx, M.J., *Spatial on-line analytical processing (SOLAP): concepts, architectures and solutions from a geomatics engineering perspective*, In Koncillia, W.R. (Eds.), *Data warehouses and OLAP: concepts, architectures and solutions*, 2005.
- Bédard, Y. and Vallière, D., *Qualité des données à référence spatiale dans un contexte gouvernemental*, Université Laval, Québec, 1995. 53 p.
- Bejaoui, L., *Qualitative topological relationships for objects with possibly vague shapes: implications on the specification of topological integrity constraints in transactional spatial databases and in spatial data warehouses*, Ph.D. Dissertation, Université Laval, 2009. 227 p.
- Benslimane, D., *Interopérabilité de SIG : la solution Isis*, *Revue internationale de géomatique*, 11(1), 2001. p 7-42.
- Bernier, E., P. Gosselin, T. Badard, Y. Bédard, *Easier Surveillance Of Climate-Related Health Vulnerabilities Through A Web-Based Spatial Olap Application*, *International Journal of Health Geographics* 8 (18), 2009.
- Bilodeau, J.M., *Analyse la distribution spatiale d'une population pour la planification d'échantillonnage appliquée à l'inventaire forestier*, MS Thesis, Department of Geomatics Sciences, Laval University, Quebec City, Canada, 1991. 142p.
- Bishr, Y., *Semantics Aspects of Interoperable GIS*, Ph.D. Dissertation, ITC Publication, 1997.
- Bishr, Y., *Overcoming the semantic and other barriers to GIS interoperability*, *Int. J. Geographical Information science* 12, 1998. p 299-314.
- Blane, X., *MDA en action*, Eyrolles (Eds.), Eyrolles, 2005.
- Boin, A.T., *Exposing Uncertainty: Communicating spatial data quality via the Internet*, Ph.D. Dissertation, University of Melbourne, 2008. 197p.
- Bouquet, P., Ehrig, M., Euzenat, J., Franconi, E., Hitzler, P., Krötzsch, M., Serafini, L., Stamou, G., Sure, Y. And Tessaris, S., *Specification of a common framework for characterizing alignment*, Deliverable D2.2.1 (version 2), KnowledgeWeb Network of Excellence, 2005.
<http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D2.2.1v2.pdf>
- Boussaid, O., Tanasescu, A., Bentayeb, F. and Darmont, J., *Integration and dimensionalm modelling approaches for complex data warehousing*, *Journal of Global Optimization*, 3, 2007. p 571-591

- Brassel, K., Bucher, F., Stephan, E.M. and Vckovski, A., *Completeness*, In: Guptill, S.C. and Morrison, J.L. (Eds.), *Elements of spatial data quality*, Elsevier Science inc., New York, 1995. p 81-107.
- Breitbart, Y., *Multidatabase Interoperability*, SIGMOD RECORD, 19 (3), 1990.
- Brézillon, P. and Pomerol, J.C., *Modelling and using context for system development: Lessons from experience*, Journal of Decision Systems, 10, 2001. p 265-288.
- Brodeur, J., *Interopérabilité des données géospatiales: élaboration du concept de proximité géosémantique*, Ph.D. Dissertation, Université Laval, 2004. 247 p.
- Brodeur, J. and Bédard, Y., *Geosemantic Proximity, a Component of spatial Data interoperability*, In: Internat. Workshop, Semantics of Enterprise Integration, ACM Conference on OOPSLA, 2001. p 14-18.
- Brodeur, J., Bédard, Y., Moulin, B., and Edwards, G., *Revisiting the concept of geospatial data interoperability with the scope of a human communication process*, Transactions in GIS, 7, 2003. p. 243-265.
- Brodeur, J. and Danko, D., *NAP – Metadata Profile*, ISO TC211, 22nd Plenary Meeting Standards in Action Workshop, Orlando, Florida, May 24, 2006.
- Brodie, M.L., *The promise of distributed computing and the challenge of legacy information systems*, In: Proceedings of the IFIP WG2'6 Database Semantics Conference on Interoperable Database Systems, 1992.
- Bruckner, R., Wang Ling, T., Mangisengi, O., and Tjoa, A.M., *A framework for a multidimensional OLAP model using topic maps*, In: Claramunt, C. et al. (Eds.), *Second International Conference on Web Information Systems Engineering (WISE'01)*, Dec. 2001 (IEEE Computer Society, Kyoto, 2001). p 109-18.
- Buehler, K. and McKee, L. (Eds.), *The Open GIS Guide (Part I): Introduction to Interoperable Geoprocessing*, Wayland, Mass: Open GIS Consortium Inc, 1996. <http://www.ogis.org/guide/guide1.htm>.
- Cai, G., Wang, H. and MacEachren, AM., *Communicating Vague Spatial Concepts in Human-GIS Interactions: A Collaborative Dialogue Approach*, Lecture Notes in Computer Science, 2003. p 287-300.
- Castano, S., Ferrara, A. and Montanelli, S., *Ontology-based Interoperability Services for Semantic Collaboration in Open Networked Systems*, 1st International Conference on Interoperability of Enterprise S/W and Applications, 2005.
- Cercone, N., Morgestern, M., Sheth, A. and Litwin, W., *Resolving semantic heterogeneity*, Panel at the International Conference on Data Engineering, 1990.
- Chamberlin, D., *XQuery: an XML query language*, IBM Systems Journal, 41(4), 2002. p 597-615.
- Chatterjee, A. and Segev, A., *Resolving Data Heterogeneity in Scientific Statistical Databases*, In: Scientific and Statistical Database Management (Proc. 6th SSDBM), 1992. p 145-159.
- Chaudhuri, S. and Dayal, U., *Data Warehousing and OLAP for Decision Support*, ACM SIGMOD Record, 26(2), 1997. p 507-508.
- Cherfi, SS., Akoka, J. and Comyn-Wattiau, I., *Perceived vs. measured quality of conceptual schemas: an experimental comparison*, ACM International Conference Proceeding Series, 334, 2007.
- Choi, K. and Luk, W., *Processing Aggregate Queries on Spatial OLAP Data*, Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery, Turin, Italy, 2008. p 125-134.

- Chorley, R.J. and Haggett, P., *Models in Geography*, Barnes & Noble, 1967. 816 p.
- Chrisman, N.R., *The Role of Quality Information in the LongTerm Functioning of a Geographical Information System*, Proceedings of International Symposium on Automated Cartography (Auto Carto 6), Ottawa, Canada, 1983. p 303-321
- Chrisman N., *Exploring geographic information systems*, John Wiley & Sons, New York, 1997. 298 p.
- Codd, E.F., Codd, S.B. and Salley, C.T., *Providing OLAP (On-Line Analytical Processing) to User- Analysts: An IT Mandate*, Hyperion white papers, 1993. 20 p.
- Colliat, G., *OLAP, Relational and Multidimensional Database Systems*, ACM SIGMOD Record, 25(3), 1995. p 64-69.
- Corey, M. and Abbey, M., *Oracle Data Warehousing*, Osborne Publishing, 1996. 384 p.
- Craik, K., *The Nature of Explanation*, Cambridge University Press, Cambridge 1943.
- Daconta, M.C., Obrst, L. and Smith, K.T., *The Semantic Web: A guide to the future of XML, Web Services and Knowledge Management*, New York, Wiley, 2003. 145p.
- Damiani, M.L. and Spaccapietra, S., *Spatial Data Warehouse Modeling. Processing and Managing Complex Data for Decision Support*, Darmont, J. and Boussaid, O. (Eds.), Idea Group Publishing, 2006. p 1-27.
- Darmont, J., Boussaid, O., Ralaivao, J., Aouiche, K., An architecture framework for complex data warehouses. 7th International Conference on Enterprise Information Systems (ICEIS 05), 2005. p. 370-373.
- Dassonville, L., Vauglin, F., Jakobsson A. and Luzwt, C., *Quality Management, Data Quality and Users, Metadata for Geographical Information*, In Spatial Data Quality, edited by Shi, W., Fisher, P.F. and Goodchild, M.F. (Taylor & Francis), 2002. p 202-215.
- David, B. and Fasquel, P., *Qualité d'une base de données géographique: concepts et terminologie*, Bulletin d'information de l'IGN, 67, IGN France, 1997.
- De Bruin, S., Bregt, A. and Van de Ven, M., *Assessing fitness for use: the expected value of spatial data sets*, International Journal of Geographical Information Systems, 15, 2001. p 457-471.
- Denes, P., and Pinson, E., *The speech chain. Physics and biology of spoken language*, London: W. H. Freeman, 1993.
- Denk, M. and Oropallo, F., *Overview of the Issues in Longitudinal and Cross-Sectional Multi-Source Databases*. Work Package No. 1: Integration of cross-section and longitudinal microdata from surveys and/or administrative registers: statistical issues, 2002.
- Derby, S.L. and Keeney, R.L., *Risk analysis: understanding how safe is safe enough*, Risk Analysis, 1(3), 1981. 217-224 p.
- Devillers, R., *Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales*, Ph.D. Dissertation, Université Laval, 2004. 167p.
- Devillers, R., Bédard, Y., Jeansoulin, R. and Moulin, B., *Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data*, International Journal of Geographical Information Science, 21(3), 2007. p 261-282.
- Devillers, R., Gervais, M., Bedard, Y. and Jeansoulin, R., *Spatial data quality: From metadata to quality indicators and contextual end-user manual*, In Paper presented at OEEPE/ISPRS joint workshop on Spatial Data Quality Management, Istanbul, March 2002.

- Dobson J.E., *The GIS revolution in science and society*, in Geography and Technology (Eds.) S D Brunn, S L Cutter, J W Harrington Jr (Kluwer, Dordrecht), 2004. p 573-587.
- Dubé, E., Badard, T. and Bédard, Y., *XML Encoding and Web Services for Spatial OLAP Data Cube Exchange: an SOA Approach*, Journal of Computing and Information Technology, 17 (1), 2009. 10p.
- Dumedah, G., *Exploring the characterization of uncertainty in census and borehole data using rough sets*, Master's thesis, Simon Fraser University, 2005.
- Durbha, S.S., King, R.L., Shah, V.P. and Younan, N.H., *A framework for semantic reconciliation of disparate earth observation thematic data*, J. Computers&Geosciences, 35, 2009. p 761-773.
- Eckert, K., Meilicke, C. and Stuckenschmidt, H., *Improving Ontology Matching Using Meta-level Learning*, The Semantic Web: Research and Application, Springer, 5554, 2009.
- Edwards, G., *A virtual test bed in support of cognitively-aware geomatics technologies*, Conference on Spatial Information Theory (COSIT). Lecture Notes in Computer Science 2205, Springer, 2001. p 140-155.
- Egenhofer, M. J., *A model for detailed binary topological relationships*, Geomatica, 47, 1993. p 261-273.
- Egenhofer, M. J. and Franzosa, R. D., *Point-set topological spatial relations*, International Journal of Geographic Information Science, 5, 1991. p 161-174.
- Eklöf, M., Mårtensson, C., *Ontological Interoperability*, Totalförsvarets, 2006.
- Few, S., *Information Dashboard Design: The Effective Visual Communication of Data*, Sebastopol, CA: O'Reilly Media, 2006.
- Fileto, R., *Issues on interoperability and integration of heterogeneous geographical data*, In III Brazilian Symposium on Geoinformatics, GEOINFO, Rio de Janeiro, Brazil, 2001. p 133-140.
- Fonseca, F. T., Egenhofer, M. J., Davis, C. A. and Câmara, G., *Semantic granularity in ontology-driven geographic information systems*, Annals of Mathematics and Artificial Intelligence, 36(1-2), 2002 (a). p 121-151.
- Fonseca, F. T., Egenhofer, M. J., Agouris, P. and Câmara, G., *Using Ontologies for Integrated Geographic Information Systems*, Transactions in GIS, 6(3), 2002 (b). p 231-257.
- Frank, A.U., *Building a spatial data framework – finding the best available data*, *Data Quality in Geographic Information: from error to uncertainty*, Hermès, Paris, 1998. 192p.
- Frank, A.U., Grum, E. and Vasseur, B., *Procedure to select the best dataset for a task*, Proceedings of GIScience, Adelphi, USA, 2004. p 81-93.
- Frank A.U., *Data Quality Ontology: An Ontology for Imperfect Knowledge*, COSIT 2007, LNCS 4736, 2007. p 406-420.
- Frank, S.C. and Chen, C.W., *Integrating Heterogeneous Data Warehouses Using XML Technologies*, Journal of Information Science, 31 (3), 2005. p 209-229.
- Franklin, C., *An Introduction to Geographic Information Systems: Linking Maps to Databases*, Database, 15(2), 1992. p 13-21.
- Gaumont, F., *Analyse du potentiel d'interopérabilité sémantique et spatiale des données d'inventaire forestier issues d'entités administratives différentes*, Master's thesis, Université Laval, 1998. 104 p.

- Gazzaniga, M.S., *Mind Matters: How Mind and Brain Interact to Create Our Conscious Lives*, Houghton Mifflin Co., Boston, 1988. 255p.
- Genero, M., Manso, E., Visaggio, A., Canfora, G. and Piattini, M., *Building measure-based prediction models for UML class diagram maintainability*, Empir Software Eng, 12, 2007. p 517-549.
- Genesereth, M.R. and Nilsson, N.J., *Logical Foundations of Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann Publishers, 1987.
- Gervais, M., *Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques*, Thèse de doctorat, Université Laval et Université Marne-La-Vallée, 2004.
- Giunchiglia, F., McNeill, F., Yatskevich, M., Pane, J., Besana, P., and Shvaiko, P., *Approximate structure-preserving semantic matching*, Proceedings 7th Conference on Ontologies, Databases, and Applications of Semantics (ODBASE), Monterey, Mexico, November-13 2008. p 1234-1253.
- Giunchiglia, F., and Yatskevich, M., *Element Level Semantic Matching*, In Proceedings of Meaning Coordination and Negotiation Workshop at International Semantic Web Conference, 2004.
- Goh, C.H., *Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Systems*, Ph.D. Dissertation, Massachusetts Institute of Technology, 1997.
- Gomez-Perez, A., *Ontological engineering: A state of the art*, Expert Update, 2 (3), 1999. p 33-43.
- Goodchild, M.F. and Gopal, S., *Accuracy of Spatial Databases*, Taylor and Francis: London, 1990. 308 p.
- Goodchild, M.F., Egenhofer, M.J. Fegeas, R., and Kottman, C., *Interoperating Geographic Information System*, Boston, Massachusetts, Kluwer Academic Publishers, 1999.
- Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F. and Pirahesh, H., *Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub Totals*, Data Mining and Knowledge Discovery, 1(1), 1997. p 29-53.
- Gruber, T.R., *Toward principles for the design of ontologies used for knowledge sharing*, Original paper presented at the International Workshop on Formal Ontology, 1993.
- Gruninger, M. and Kopena, J., *Semantic integration through invariants*, In Doan, A., Halevy, A. and Noy, N. (Eds.), Workshop on Semantic Integration at ISWC-2003, Sanibel Island, FL, 2003.
- Guarino, N., *Formal Ontologies and Information Systems*, In: Guarino, N. (Eds.), Proceedings of of the International Conference on Formal Ontology and Information Systems (FOIS'98), IOS Press, 1998. p 3-15.
- Guptill, S.C. and Morrison, J.L., *Spatial data quality*, Elements of spatial data quality, Guptill, S. C. and Morrison, J. L. (Eds.), Elsevier Science inc., New York, 1995.
- Hakimpour, F., *Using Ontologies to Resolve Semantic Heterogeneity for Integrating Spatial Database Schemata*, Ph.D. Dissertation, Zürich University, 2003.
- Hammer, M. and McLeod, D., *On Database Management System Architecture*, in Infotech State of the Art Report, 8, Data Design, Pergamon Infotech Limited, 1980.
- Han, J., Stefanovic, N. and Koperski, K., *Selective Materialization: An Efficient Method for Spatial Data Cube Construction*, Proceedings Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, Australia, 1998. p 144-158.

- Hartmann, S. and Link, S., *Weak Functional Dependencies: Full Propositional Expressiveness for the Database Practitioner*, Journal of Universal Computer Science, 2009.
- Harvey, F. J., *Semantic interoperability and citizen/government interaction*, In Proceedings of spatial data handling 2002: joint international symposium on geospatial theory, processing, and applications, 2002. 13p.
- Harvey, F., Kuhn, W., Pundt, H., Bishr, Y. and Riedemann, C., *Semantic Interoperability: A Central Issue for Sharing Geographic Information*, Annals of Regional Science, Special Issue on Geo-spatial Data Sharing and Standardization, 1999. p 213-232.
- Hess, G.N., Iochpe, C. and Castano, S., *An Algorithm and Implementation for GeoOntologies Integration*, Proceedings 8th Symposium on GeoInformatics, Campos do Jordao, Brasil, 2006. p 129-140.
- Hümmer, W., Bauer, A. and Harde, G., *XCube-XML for data warehouses*, In: Proceedings of the ACM 6th International Workshop on Data Warehousing and OLAP, DOLAP'03, ACM, Louisiana, 2003. p 33-40.
- Hunter, G.J. and Reinke, K.J., *Adapting Spatial Databases to Reduce Information Misuse Through Illogical Operations*, Lemmens, M. and Heuvelink, G. (Eds.), Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Amsterdam, 2000. p 313-319.
- IEC 61508. International Standard: Functional Safety of Electrical/Electronic/Programmable Electronic Systems. International Electrotechnical Commission, Geneva, 2000.
- ISO 31000. Risk management – Principles and guidelines, International Standard ISO 31000:2009, International Organization for Standardization, Geneva, 2009.
- ISO JTC 1/SC 34. DTR 13250-1: Information technology -- Topic Maps -- Part 1: Overview and basic concepts, Stage: 30.99, 2008.
- ISO/TC 211 19103. Geographic Information - Conceptual Schema Language. Geneva, Switzerland, International Organization for Standardization, 2005 (a).
- ISO/TC 211 ISO/DIS 19110. Geographic Information - Feature Cataloguing Methodology. Geneva, Switzerland, International Organization for Standardization, 2005 (b).
- ISO/TC 211 19113. Geographic information - Quality principles, International Organization for Standardization (ISO), 2002.
- ISO/TC 211 19115. Geographic information - Metadata. International Organization for Standardization, 2003.
- Janowicz, K., *Service Composition and Interoperation with DAML-S: A critical view*, Diploma thesis, University of Münster, 2003.
- Janowicz, K., Wilkes, M. and Lutz, M., *Similarity-Based Information Retrieval and Its Role within Spatial Data Infrastructures*, Geographic Information Science, 5266, 2008. p 151-167.
- Java™ API for XML Processing (JAXP). Web Site: <http://java.sun.com/webservices/reference/tutorials/jaxp/html/docinfo.html> (last visited: March 2010)
- Juran, J.M., Gryna, F.M.J. and Bingham, R.S., *Quality Control Handbook*, New-York, McGraw-Hill, 3rd Edition, 1979.
- Kahn, B.K. and Strong, D.M., *Product and Service Performance Model for Information Quality: An Update*, Proceedings of Conference on Information Quality, Cambridge, USA, 1998. p 102-115.

- Kashyap, V. and Sheth, A., *Semantic Heterogeneity in Global Information System: The Role of Metadata, Context and Ontologies*, in Cooperative Information Systems: Current Trends and Directions, Papazoglou, M. and Schlageter, G. (Eds.), London, Academic Press, 1996. p 139-178.
- Kearns, J.T., *Thinking Machines: Some Fundamental Confusions*, Minds and Machines, 7, 1997. p 269-287.
- Klug, A., *Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions*, Journal of the ACM, 29(3), 1982. p 699-717.
- Krishnamurthy, R., Litwin, W. and Kent, W., *Language features for interoperability between databases with schematic discrepancies*, In: Clifford, J. and King, R. (Eds.), Proceedings of the ACM SIGMOD-International Conference on Management of Data ,ACM, Denver, 1991. p 40-49.
- Kuhn, W., *Geogeospatial Semantics: Why, of What, and How*, Journal on Data Semantics, 2, 2005.
- Kwon, O.B., Choi, S.C. and Park, G.R., *NAMA: a context-aware multi-agent based web service approach to proactive need identification for personalized reminder systems*, Expert Systems with Applications, 29, 2005. p 17-32.
- Levesque, M.A., Bédard, Y., Gervais, M. and Devillers, R., *Towards managing the risks of data misuse for geospatial datacubes de données*, 5th International Symposium on Spatial Data Quality , ISSDQ 2007, Enchede, Pays-Bas, 2007.
- Lutz, M. and Klien, E., *Ontology-based retrieval of geographic information*, International Journal of Geographical Information Science, 20(3), 2006. p 233-260.
- MacEachren A.M., *How maps work: representation, Visualization and Design*, The Guilford Press, New York, 1995, 513p.
- MacEachren, A.M. and Kraak, M. J., *Exploratory cartographic visualization: advancing the agenda*, Computers & geosciences, 1997, 23(4). p 335-343.
- Machlis, G. E. and ERosa, A., *Desired risk: Broadening the social amplification of risk framework*. Risk Analysis, 10, 1990. p 161-68.
- Maedche, A. and Staab, S., *Ontology Learning for the Semantic Web*, IEEE Intelligent Systems, 16(2), 2001. p 72-79.
- Malinowski, E., Zimányi, E., *Representing spatiality in a conceptual multidimensional model*, In: Proceedings of ACM int. workshop on Geographic information systems, ACM Press, New York, 2004.
- Malinowski, E. and Zimányi, E., *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*, Springer, 2008. 444 p.
- Mangisengi, O., Huber, J., Hawel, C. and Essmayr, W., *A framework for supporting interoperability of data warehouse islands using XML*, hird Int'l Conf. Data Warehousing and Knowledge Discovery (DaWaK'01), 2001. p 328-38.
- Mennis, J. L., Peuquet, D. J. et Qian, L., *A conceptual framework for incorporating cognitive principles into geographical database representation*, International Journal of Geographic Information Science, 14(6), 2000. p 501-520.
- Miller, J. and Mukerji, J., *MDA Guide Version 1.0.1*, Object Management Group, Framingham, Massachusetts, 2003.
- Moe, N.B. and Smite, D., *Understanding Lacking Trust in Global Software Teams: A Multi-Case Stu*”. In: J. Münch and P. Abrahamsson (eds.) Product-Focused Software Process Improvement, LNCS (4589), 2007. p 20-32.

- Molenaar, M., *A syntax for the representation of fuzzy spatial objects*, In Proceedings of conference on Spatial data Modeling and Query Languages for 2D and 3D Applications, edited by Molenaar, M. and de Hoop, S. (Delft, NL: Geodesy-New Series), 1994. p 155-169.
- Moody, D. L., *Theoretical and Practical Issues in Evaluating the Quality of Conceptual Models: Current State and Future Directions*, Data & Knowledge Engineering, 15(3), 2005. p 243-276.
- Morgan, M. G., *Choosing and Managing Technology-Induced Risks*, Readings in Risk. In Glickman T.S. and Gough M. (Eds.). Resources for the Future, Washington, 1990. p 5-15.
- Nagarajan, M., Verma, K., Amit, P. S., Miller, J. and Lathem, J., *Semantic Interoperability of Web Services*, Challenges and Experiences, 2006.
- Nambiar, U., Ludäscher, L., Lin, K. and Baru, C., *The GEON portal: accelerating knowledge discovery in the geosciences*, Proceedings 8th ACM International Workshop on Web Information and Data Management (WIDM), Arlington, Virginia, USA, November 10, 2006. p 83-90.
- Nienaber, R. C., *A model for enhancing software project management using software agent technology*, Diploma thesis, University of South Africa, 2008
- Norman, D., *Some observations on mental models*. In Gentner, D. and Stevens, A.L. (Eds.), Mental Models. Hillsdale, N.J.:Lawrence Erlbaum Associates, 1983. p 7-14.
- Nwana, H. S., *Software agents: an overview*, The Knowledge Engineering Review, 11, 1996. p 205-244.
- Nwana, H. S. and Wooldridge, M., *Software agent technologies*, British Telecommunications Technology Journal 14 (4), 1996. p 68-78.
- Nyerges T., *Representing geographical meaning*, In Buttenfield, B.P. and McMaster, R.B. (Eds.), Map Generalization: Making Rules for Knowledge Representation, 1991. p 59-85.
- Obrst, L., *Ontologies for Semantically Interoperable Systems*, CIKM, 2003. p 366-369.
- Ogden C.K. and Richards, I.A., *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, Routledge & Kegan Paul Ltd., London, 1923.
- OLAP Council, *OLAP and OLAP Server Definitions*, 1995. <http://www.olapcouncil.org/research/glossary.htm>.
- OMG-IBM, *Ontology Definition Metamodel*, Third Revised Submission to OMG/ RFP ad/2003-03-40, 2003.
- Ouksel, A. and Sheth, A., *Semantic Interoperability in Global Information Systems: A Brief Introduction to the Research Area*, SIGMOD Record, 28(1), 1999. p 5-12.
- Papadias, D., Tao, Y., Kalnis, P. and Zhang, J., *Indexing Spatio-temporal Data Warehouses*, IEEE Data Engineering Bulletin, 25(1), 2002. p 10-17.
- Park, J. and Ram, S., *Information Systems Interoperability: What Lies Beneath?*, ACM Transactions on Information Systems, 2004.
- Payne, T. R., Paolucci, M., Singh, R. and Sycara, K., *Communicating agents in open multi agent systems*, In Proceedings of first GSFC/JPL workshop on radical agent concepts (WRAC), 2002. p1-10.
- Pedersen, T.B. and Jensen, C.S., *Multidimensional Database Technology*, IEEE Computer, 34(12), 2001. p 40-46.

- Pedersen, D., Riis, K. and Pedersen, T.B., *XML-Extended OLAP Querying*, Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM'02), IEEE Computer Society Press, 2002. p 195-206.
- Pendse, N., Glossary, The OLAP Report, 2000. <http://www.olapreport.com/glossary.htm>. Accessed October 12, 2005.
- Pepper, S. and Moore, G., *XML Topic Maps (XTM) 1.0 TopicMaps*, Org Specification, 2001. (Available at: www.topicmaps.org/xtm/1.0/).
- Pipino, L. L., Lee, Y.W. and Wang, R.Y., *Data quality assessment*, Communications of ACM, 45, april, 2002. p 211-218.
- Poore, B. and Chrisman, N., *Order from noise: Toward a social theory of geographic information*. Annals of the Association of American Geographers, 96(3), 2006. p 508-523.
- Pourabbas, E. and Rafanelli, M., *Characterization of Hierarchies and Some Operators in OLAP Environment*, Proceedings of the 2nd ACM International Workshop on Data Warehousing and OLAP, Kansas City, United States, 1999. p 54-59.
- Rafanelli, M., *Multidimensional Databases: Problems and Solutions*, Idea Group Publishing, 2003. 473p.
- Reenskaug T., Wold, P. and Lehne, O.A., *Working with objects: The OOram Software Engineering Method*, Taskon, Gaustadalléen 21, N-0371 Oslo 3 Norway, 1995.
- Renn, O., *Three decades of risk research: accomplishments and new challenges*, Journal of Risk Research, 1(1), 1998. p 49-71.
- Rhyne, T.M., MacEachren, A. and Dykes, J., *Exploring Geovisualization*, IEEE Computer Graphics and Applications, 2006. p 20-21.
- Rifaieh, R.D., *Utilisation des ontologies contextuelles pour le partage sémantique entre les systèmes d'information dans l'entreprise*, Ph.D. Dissertation, Institut National des sciences Appliquées de Lyon, 2004.
- Rivest, S., Bédard, Y. and Marchand, P., *Towards better support for spatial decision-making: Defining the characteristics of Spatial On-Line Analytical Processing (SOLAP)*, Geomatica, the journal of the Canadian Institute of Geomatics, 2001. p 539-555.
- Rivest, S., Bédard, Y., Proulx, M., Nadeau, M., Hubert, F. and Pastor, J., *SOLAP Technology: Merging Business Intelligence with Spatial Technology for Interactive Spatio-Temporal Exploration and Analysis of Data*, J. ISPRS Advances in spatio-temporal analysis and representation, 2005. p 17-33.
- Rodriguez, M.A., *Assessing Semantic Similarity Among Entity Classes*, Ph.D. Thesis, University of Maine, 2000.
- Roehrig, J., *Information Interoperability for River Basin Management*, J. Technology Resource Management and Development - Scientific Contributions for Sustainable Development, 2, 2002.
- Rosch, E., *Principles of categorization*, In: Rosch, E. and Lloyd, B. (Eds.), Cognition and Categorization, 1978. p 27-77.
- Salehi, M., *Developing a model and a language to identify and specify the integrity constraints in spatial datacubes*, Ph.D. Dissertation, Université Laval, 2009 . 191p.
- Sboui, T., and Bédard, Y., MGsP: Extending the GsP to Support Semantic Interoperability between Geospatial Datacubes, Proceedings of ER/2010 SeCoGIS international workshop, LNCS, Springer, 2010 (a) (to be published).

- Sboui, T., Bédard, Y., Brodeur, J. and Badard, T., *Managing the Risk related to the Simultaneous Use of Spatial Datacubes*, 13th International Convention and Fair Informatica 09, Havane, Cuba. 2009 (a).
- Sboui, T., Bédard, Y., Brodeur, J. and Badard, T., *Modeling the External Quality of Context to Fine-tune Context Reasoning in Geospatial Interoperability*, The 21st Int. Joint Conference on Artificial Intelligence - Proceedings of the Workshop on Automated Reasoning about Context and Ontology Evolution (ARCOE'09) Pasadena, California, USA, 2009 (b). p 28-30.
- Sboui, T., Bédard, Y., Brodeur, J. and Badard, T., *A Conceptual Framework to Support Semantic Interoperability of Geospatial Datacubes*, Proceedings of ER/2007 SeCoGIS workshop, LNCS, Springer, 2007. p 378-387.
- Sboui, T., Salehi, M., and Bédard, Y., *A systematic approach for managing the risk related to semantic interoperability between geospatial datacubes*, International Journal of Agricultural and Environmental Information Systems, 2010 (b).
- Sboui, T., M. Salehi, Bédard, Y., and Rivest, S., *Catégorisation des problèmes d'intégration des modèles des cubes de données spatiales*, Extraction et Gestion des Connaissances conference, Sophia Antipolis, France, 2008.
- Sboui, T., Salehi, M., and Bédard, Y., *Towards a Quantitative Evaluation of Geospatial Metadata Quality in the Context of Semantic Interoperability*, ISSDQ 2009, July 5-8th, St-John's New Founland, Canada, 2009 (c).
- Sboui, T., M. Salehi, Bédard, Y., and Rivest, S., *Une approche basée sur la qualité pour faciliter l'intégration de modèles de cubes de données spatiales*, Revue des Nouvelles Technologies de l'information, 2009 (d).
- Sboui, T., Yaagoubi, R., Bédard, Y., and Edwards, G., *SemEL: A Semantic Model Informed by Cognitive Principles to Support Reasoning about Spatial Data Semantics*, Las Navas 2010: Cognitive and Linguistic Aspects of Geographic Space, Las Navas, Spain, 2010 (c).
- Schramm W., *The Nature of Communication Between Humans*, In Schramm, W. and Robert, D.F. (Eds.), *The Process and Effects of Mass Communication*, Champaign-Urbana, IL, University of Illinois Press, 1971. p 3-53.
- Schmidt A., Beigl, M. and Gellersen, H. W., *There is more to context than location*, Computers and Graphics, 23 (6), 1999. p 893-901.
- Sciore, E., Siegel, M. and Rosenthal, A., *Using semantic values to facilitate interoperability between heterogeneous information systems*, ACM Transactions on Database Systems, 19(2), 1994 a. p 254-90.
- Sciore, E., Siegel, M. and Rosenthal, A., *Context interchange using meta-attributes*, ACM Transactions on Database Systems, 19(2), 1994 b. p 254 -290.
- Semy, S., Pulvermacher, M. and Obrst, L., *Toward the Use of an Upper Ontology for U.S. Government and Military Domains: An Evaluation*, MITRE Technical Report 04B0000063, September 2004.
- Sester, M., *Knowledge acquisition for the automatic interpretation of spatial data*, International Journal of Geographic Information Science, 14 (1), 2000. p 1-24.
- Shannon, C.E., *A Mathematical Theory of Communication*, The Bell System Technical Journal, 27, 1948. p 379-423, p 623-656.
- Shekhar, S., Lu, C.T., Tan, X., Chawla, S. and Vatsavai, R., *Map Cube: A Visualization Tool for Spatial Data Warehouses*, Geographic Data Mining and Knowledge Discovery, Miller, H. J. and Han, J. (Eds.), Taylor & Francis, 2001. p 73-108.

- Sheth, A., *Changing Focus on Interoperability in Information Systems: From Systems, Syntax, Structure to Semantics*, In Goodchild, M., Egenhofer, M., Fegeas, R. and Kottman, C. (Eds.), *Interoperating Geographic Information Systems*, Boston, Massachusetts, Kluwer Academic Publisher, 1999. p 5-29.
- Sheth, A., and Kashyap, V., *So far (schematically), yet so near (semantically)*, In Proceedings of the IFIP TC2/WG2.6 Conference on Semantics of Interoperable Database Systems, DS-5, In IFIP Transactions A-25, North, 1992.
- Sheth, A. and Larson, J., *Federated database systems for managing distributed, heterogeneous, and autonomous databases*, ACM Computing Surveys, 22(3), 1990. p 183-230.
- Smith, M.K., Welty, C. and McGuinness, D.L., *OWL Web Ontology Language Guide*, 2004. (Web site: <http://www.w3.org/TR/owl-guide/>)
- SOPFEU, Société de protection des forêts contre le feu, *Rapport annuel 2008*. The report can be found at http://www.sopfeu.qc.ca/imports/_uploaded/SopfeuRA08Final.pdf
- Sowa, J.F., *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole, Pacific Grove, CA, 2000.
- Staub, P., Gnagi, H.R. and Morf, A., *Semantic Interoperability through the Definition of Conceptual Model Transformations*, Transactions in GIS, 12(2), 2008. p 193-207.
- Snyder, J.P., *Flattening the Earth: Two Thousand Years of Map Projection*, University of Chicago Press, 1997.
- Swanson, D.R. and Smalheiser, N.R., *An interactive system for finding complementary literatures: a stimulus to scientific discovery*, Artificial Intelligence, 91 (2), 1997. p 183-203.
- Sugumaran, V. and Storey, V.C., *Ontologies for Conceptual Modeling: Their Creation, Use, and Management*, Data & Knowledge Engineering, 2002. p 251-271.
- The Apache Xalan Project, *Xalan-Java version 2.7.1*, 2006. Web site: <http://xml.apache.org/xalan-j/>
- The Apache Xerces Project, *Xerces-Java Java Parser*, 2005. Web site: <http://xerces.apache.org/xerces-j/>
- The OpenGIS Abstract Specification (OGC): Topic 12: OpenGIS Service Architecture. Version 4.3, http://portal.opengeospatial.org/files/?artifact_id=1221, 2002.
- Timpf, S., Raubal, M. and Werner, K., *Experiences with Metadata*, Proceedings of Symposium on Spatial Data Handling, SDH'06, Advances in GIS Research II, Delft, The Netherlands, 1996. p 12B.31-12B.43.
- Turban, E. and Aronson, J.E., *Decision Support Systems and Intelligent Systems*, 6th Edition, Prentice Hall, 2001.
- Uitermark, H.T., van Oosterom, P.J.M., Mars, N.J.I. and Molenaar, M., *Ontology-based geographic data set integration*, In Böhlen, M. H., Jensen, C. S. and Scholl, M. O. (Eds.), *Proceedings of spatio-temporal database management, international workshop STDBM'99*, 1678. Lecture notes in computer science, Berlin, Heidelberg, Springer-Verlag, 1999. p 60-78.
- Unified Modeling Language (UML). Web site: <http://www.uml.org/> (last visited: March 2010).
- Vaccari, L., Shvaiko, P. and Marchese, M., *A geo-service semantic integration in Spatial Data Infrastructures*, International Journal of Spatial Data Infrastructures Research, 4, 2009. p 24-51.

- Van Oort, P., *Spatial data quality: from description to application*, Publication on Geodesy 60, Netherlands Geodetic Commission, ISBN 90 6132 295 2, Delft, December, 2006.
- Vieira, V., Salgado, A. and Tedesco, P., *Towards an Ontology for Context Representation in Groupware*, Proceedings of the 11th International Workshop, CRIWG, Brazil, 2005. p 367-375.
- Visser, U. and Stuckenschmidt, H., *Interoperability in GIS - Enabling Technologies*, In Proceedings of 5th AGILE Conference on Geographic Information Science, Ruiz, M., Gould, M. and Ramon, J. (Eds.), Palma de Mallorca, Spain, 2002. p 291-297.
- Voisard, A. and Schweppe, H., *Abstraction and decomposition in interoperable GIS*, International Journal of Geographical Information Science, 1998.
- Wand, Y. and Weber, R.A., *Research commentary: information systems and conceptual modelling*, a research agenda, Information Systems Research, 13 (4), 2002. p 363-376.
- Webster, K. P. B., De Oliveira, K. M. and Anquetil, N., A., *Risk Taxonomy Proposal for Software Maintenance*, Proceedings of the 21st IEEE International Conference on Software Maintenance, ICSM '05, IEEE Computer Society, 2005.
- Wikipedia: <http://www.wikipedia.com/> (Last visted March, 2010)
- Wooldridge. M., *Introduction to MultiAgent Systems*, 1st ed, John Wiley & Sons, 2002.
- Worboys, M. and Clementini, E., *Integration of imperfect spatial information*, Journal of Visual Languages and Computing, 12 (1), 2001. p 61-80.
- W3C, *Simple Object Access Protocol (SOAP) 1.2*, W3C Note. <http://www.w3c.org/TR/SOAP>. 2007.
- W3C, *Web Service Description Language (WSDL) 1.1*, W3C Note. <http://www.w3.org/TR/wsdl.html>. 2001.
- Xhu, Z. and Lee, Y.C., *Semantic Heterogeneity of Geodata*, In Proceedings of ISPRS Commission IV Symposium, 2002.
- Yougworth, P., *OLAP Spells Success For Users and Developers*, Data Based Advisor, 1995. p 38-49.
- Yu, C., Sun, W., Dao, S. and Keirse, D., *Determining relationships among attributes for Interoperability of Multidatabase Systems*, Proc. of the 1st International Workshop on Interoperability in Multidatabase Systems, Kyoto, Japan, April, 1991.
- Zhao, S., Wongthongtham, P., Chang, E. and Dillon, T., *Knowledge Conceptualization and Software Agent based Approach for OWL Modeling Issues*, Artificial Intelligence in Theory and Practice II (IFIP International Federation for Information Processing), Bramer, M. (Ed), Boston, Springer, 276, 2008. p 361-370.

Annex A: An example of generic context elements

In order to guide our search for the generic contexts, we defined a set of generic context elements based on five axes: What, Where, When, Why and How. These kinds of questions can be found in metadata. That is why we based our search for the global context elements on the metadata standards of geospatial information and geospatial metadata. We used ISO/TC 211 and FGDC standards which type the metadata elements of geospatial information.

We defined four levels, starting from the most generic level (see Table A.1). The first level is Goal Context level. In this level, we define the purpose for which the data will be used as well as some general conditions or situations in which the application will be developed. In our research project, the global context is the interoperability between geospatial datacubes. The second level, called Domain Context contains the contexts which define the domain in which the application will be developed and used. In our research project, this level contains the domains and basic concepts related to the geospatial domain (for example, we can find a cadastral application as the application domain, NAD 83 as a datum, etc). Within the third level we defined the elements related to the geospatial dataset stored in datacubes (such as the spatial domain of the dataset, DB environment (such as the used modeling technique (UML, ontology, etc.), the used SGBD, etc.), the standards used for the dataset). The fourth level includes the characteristics of dimensions, members or measures of datacubes (roles, properties, etc).

We should notice that these context elements do not aim at being complete but at providing some examples of context in the geographic domain.

<u>Context Levels</u>	<u>Context Elements</u>
4- Global context level	Spatial Context, Geospatial Multidimensional DB, Interoperability
3- Domain context level	Application domain (Keywords), Temporal reference (Correctness reference), Spatial reference (coordinate system (Type), Indirect Spatial Reference, projection system, cadastral reference system, Reference system transformation),
2- DB context level	Spatial Domain (place, location), DB theme, DB issue/subject, DB goal/intent, General data description, DB environment (Modeling technique, Development language/SGBD, Natural language, Constraint language), Data set environment Information (acquisition, instrument, Platform) , Standards, Aggregation information (method, criteria), Scale, security information (use/access constraints, source of the data set, fees) Data quality information (Precision, Lineage, Accuracy, Completeness, Consistency), Time for the DB (History, Time Period of Content), Maintenance and update frequency, Cross Reference/Related data set, Spatial representation (Format, Symbol, Resolution).
1- Concept context level	Concept properties, Concept type, Concept role, Concept relationships, Concept dimension, Concept History, Concept Constraint, Concept Values Domain, Static or dynamic, Generalization information, Related domain to which the concept can be used, Object survey.

Table A.1: An example of generic context elements.

Annex B: A Conceptual Framework to Support Semantic Interoperability of Geospatial Datacubes

Tarek Sboui¹², Yvan Bédard¹², Jean Brodeur³¹, Thierry Badard¹

¹ Department of Geomatic Sciences and Centre for Research in Geomatics, Université Laval, Quebec, Qc, G1K 7P4, Canada

² NSERC Industrial Research Chair in Geospatial Databases for Decision-Support

³ Natural Resources Canada, CIT, 2144-010 King West St, Sherbrooke, Qc, J1J 2E8, Canada

Tarek.Sboui.1@ulaval.ca, Yvan.Bedard@scg.ulaval.ca, brodeur@nrcan.gc.ca, Thierry.Badard@scg.ulaval.ca

Abstract. Today, we observe a wide use of geospatial databases that are implemented in many forms (e.g., transactional centralized systems, distributed databases, multidimensional datacubes). Among those possibilities, the multidimensional datacube is more appropriate to support interactive analysis and to guide the organization's strategic decisions, especially when different epochs and levels of information granularity are involved. However, one may need to use several geospatial multidimensional datacubes which may be heterogeneous in design or content. Overcoming the heterogeneity problems in a manner that is transparent to users has been the principal aim of interoperability for the last fifteen years. In spite of successful initiatives and widespread use of standards, today's solutions do not address yet geospatial datacubes. This paper aims at describing the interoperability of geospatial datacubes, defining the semantic heterogeneity problems that may occur when using different geospatial datacubes, and proposing a conceptual framework to support semantic interoperability of these datacubes.

A model for semantic interoperability of geospatial datacubes

We believe that, in order to overcome the semantic heterogeneity of geospatial cubes, we should reason about their semantic. We define semantic regarding the elements of ontologies (i.e., concepts, definitions, assumption, properties such as thematic, geometric, graphic and temporal aspects) and the elements of context of geospatial datacubes concepts (e.g., language, techniques used to define spatial objects, etc.). Both ontology and context elements define the semantic characteristics of geospatial data cubes concepts. In order to guide the reasoning about the concepts semantics, and inspired by the VUEL concept

(View Element) [3], we introduce a model that is based on multidimensional structure called *SemEL* (i.e., Semantic Element) where ontology and context represent the facts (see Figure B.1). This model enables to explicitly represent the meaning and to define a relevant interpretation of a concept regarding the ontology and the context in which it has been defined and used. The ontology model has five dimensions (i.e., definitions, assumptions, geometries, time and graphical representations) and a fact table that has the ontology description of datacubes concepts (*Ont_Desc*) as its unique measure. *Ont_Desc* will contain textual definition, geometry, graphical and temporal properties, as well as axioms. The context model is defined according to four dimensions (i.e., Goal Context, Domain Context, Dataset Context, and Concept Context) and a fact table that has the description of context (*Context_Desc*) as its unique measure.

Since it is based on multidimensional structure, *SemEL* enables to rapidly navigate from one level to another and from dimension to another and apply reasoning capabilities (e.g., inference) to draw conclusions based on relations between semantic elements (i.e., ontology and context elements). For example, if the term *Forest* was used in England's royal context, then by inference, this term can be interpreted as a "hunting ground". More specifically, the model would allow to:

- provide the appropriate meaning of a concept (i.e., the concept defined in the specific context, represented with a specific geometry, a specific graphic, in a specific date and according a predefined assumption). For instance, the meaning of the concept *river* can be determined by 1) its definition within a general ontology: "Natural stream of water that flows in a channel" [1], and a general assumption specifying that it flows into the sea, 2) its geometry: ^m, 3) its graphic: *blue* and 4) its context elements: *English* as the language used, agriculture as the domain in which the concept is used, etc. Consequently, the appropriate meaning of the concept *river* in *French* would be *fleuve*.
- facilitate the conversion of concepts semantics. That is, navigating through different levels of dimensions, we can change the semantic characteristics of each concept and define the impact of that change on the interpretation of this concept.
- analyze phenomena changes by facilitating the comparison of different semantic elements of the same phenomenon. In fact, *SemEL* helps to rapidly navigate through different

dimensions and compare different measures of a given phenomenon (i.e., *Ont_Desc* and *Context_Desc*) and infer what changes have affected that phenomenon. For example, if an assumption, specifying that “people can easily walk”, was added to the semantic of the concept *forest*, we can conclude that forest management was carried out.

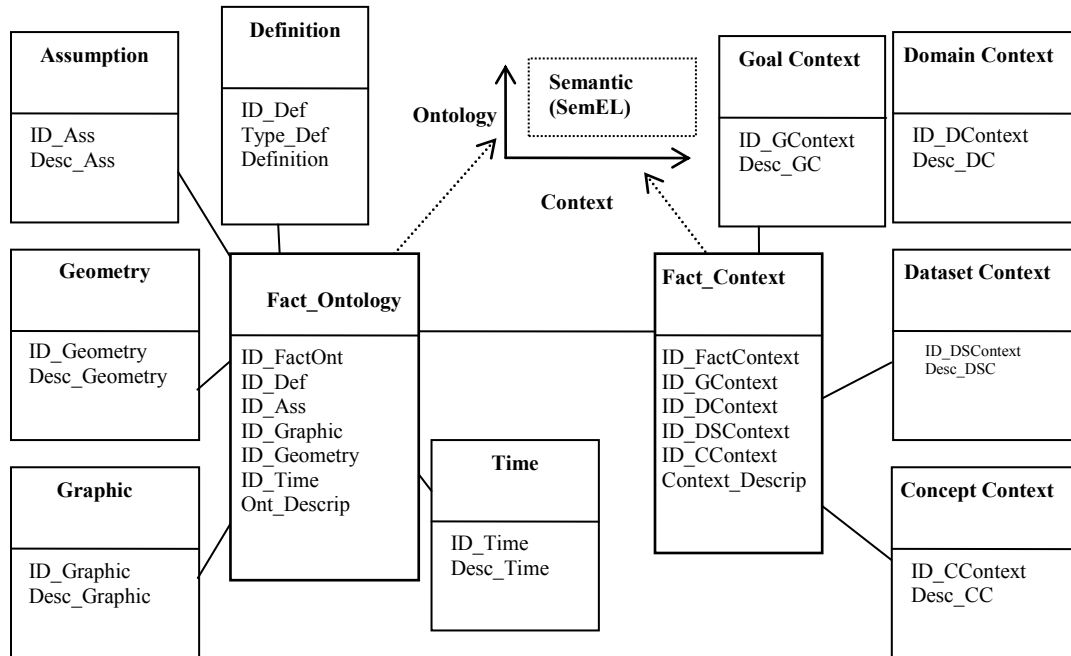


Figure B.1: A representation of the Semantic Element concept (*SemEL*).

4 Conclusion

In this paper we described the interoperability of geospatial datacubes, and we proposed a conceptual framework to overcome semantic heterogeneity problems when using geospatial datacubes. The framework is based on human communication, ontology, context, and the multidimensional structure. We defined a communication model which is based on agents representing geospatial datacubes (called *Datacubes Agents*) and a mediator agent (called *Context Agent*). The *Context Agent* helps the datacubes agents to appropriately interpret information exchanged between them. The interpretation is supported by a model which is based on the multidimensional paradigm (i.e., *SemEL*). The semantic of concepts will be

discussed regarding the dimensions of this model (i.e., elements of both ontology and context).

Further work is required to refine *SemEL* and define a mapping between two different multidimensional models. Then we would implement our framework.

Acknowledgments. We wish to acknowledge the contribution of the NSERC Industrial Research Chair in Geospatial Databases for Decision Support.

References

1. Answers. <http://www.answers.com/> (Last visted September 20, 2006)
2. Bédard, Y., Rivest, S., Proulx, M.J.: Spatial on-line analytical processing (SOLAP): concepts, architectures and solutions from a geomatics engineering perspective. In W. R. Koncillia C (Eds.) Data warehouses and OLAP: concepts, architectures and solutions (2005)
3. Bédard, Y., Bernier, E.: Supporting Multiple Representations with Spatial View Management and the Concept of "VUEL". Joint Workshop on Multi-Scale Representations of Spatial Data, ISPRS WG IV/3, ICA Com. on Map Generalization (2002)
4. Bédard, Y., Merrett, T., Han, J.: Fundamentals of spatial data warehousing for geographic knowledge discovery. H. J. Miller, J. Han (Editors) Geographic Data Mining and Knowledge Discovery (2001)
5. Bishr, Y.: Overcoming the semantic and other barriers to GIS interoperability. Int. J. Geographical Information science. 12, (1998) 299- 314
6. Brodeur, J.: Interopérabilité des données géospatiales: élaboration du concept de proximité géosémantique. Ph.D. Dissertation. Université Laval (2004)
7. Brodeur, J., Bédard, Y.: Geosemantic Proximity, a Component of geospatial Data Interoperability, Internat. Workshop, Semantics of Enterprise Integration, ACM Conference on OOPSLA (2001) 14-18

8. Brodie, M.L.: The promise of distributed computing and the challenge of legacy Information systems. In Proceedings of the IFIP WG2'6 Database Semantics Conference on Interoperable Database Systems (1992)
9. Codd, E.F., Codd S.B. Salley C.T.: Providing OLAP (On-Line Analytical Processing) to User- Analysts: An IT Mandate. Hyperion white papers, 20 p (1993)
10. Goodchild, M.F., Egenhofer, M.J., Fegeas, R.: Interoperating GIS. Report of a specialist meeting held under the auspices of the Varenus project panel on computational implementations of geographic concepts. December 5-6, Santa Barbara, California 1997.
11. Guarino, N.: Formal Ontologies and Information Systems, in Guarino N. (ed.), Proe. of FOIS98, IOS Press (1998) 3- 15
12. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Original paper presented at the International Workshop on Formal Ontology, (1993).
13. Harvey, F., Kuhn, W., Pundt, H., Bishr, Y., Riedemann, C.: Semantic Interoperability: A Central Issue for Sharing Geographic Information Annals of Regional Science. Special Issue on Geo-spatial Data Sharing and Standardization (1999) 213-232
15. Malinowski, E., Zimányi, E.: Representing spatiality in a conceptual multidimensional model. Proceedings of ACM int. workshop on Geographic information systems (2004).

Annex C: Modeling the External Quality of Context to Fine-tune Context Reasoning in Geospatial Interoperability

Tarek Sboui¹², Yvan Bédard¹², Jean Brodeur³¹, Thierry Badard¹

¹ Department of Geomatic Sciences, Université Laval, Quebec, Qc, G1K 7P4, Canada

² NSERC Industrial Research Chair in Geospatial Databases for Decision-Support

³ Natural Resources Canada, CIT, 2144-010 King West St, Sherbrooke, Qc, J1J 2E8, Canada

Tarek.Sboui.1@ulaval.ca, Yvan.Bedard@scg.ulaval.ca, brodeur@nrcan.gc.ca,
Thierry.Badard@scg.ulaval.ca

Abstract

Context reasoning is the process of drawing conclusions and inferring new information from existing context information, and is considered crucial to geospatial interoperability. However, such reasoning still remains a challenge because context may be incomplete or not relevant for a specific use. Thus, evaluating and modeling the external quality (fitness-for-use) of context information can help the process of reasoning about context. In this paper, we propose an ontology-based approach to model the external quality of context information. This approach aims at fine-tuning context reasoning and hence enhancing geospatial interoperability.

1 Introduction

Over the last decades, there has been an exponential increase in the amount of geospatial data available from multiple sources. Reusing this data can significantly decrease the cost of geospatial application. In order to develop ways to enhance the reuse of available

geospatial data, significant research efforts have been carried out. Among these efforts, semantic interoperability of geospatial data has been extensively investigated [Bishr, 1998; Harvey et al., 1999; Brodeur, 2004], but it still remains a challenge in spite of all these efforts [Staub et al., 2008]. A key issue in such interoperability is reasoning about context information of geospatial data. Context information may be used in several ways to capture the semantics of an object and its relationships to other objects. We distinguish between two kinds of context: production context and use context. Production context is any information that can be specified explicitly or implicitly by a data producer (e.g., the method of data collection). On the other hand, characteristics that surround user's application (e.g., reference system, scale) are considered as use context. Context can be thematic (e.g., data acquisition method), spatial (e.g., spatial reference system used) or temporal (e.g., the time of data acquisition). Context may be incomplete or not appropriate for a specific use. This may affect the context reasoning process [Henricksen and Indulska, 2004; Bikakis et al., 2007]. The degree of completeness and appropriateness of context information can be indicated by the quality of this context. We distinguish two parts of context quality: internal quality and external quality. The internal quality of a context refers to the extent to which a producer meets specifications, that is, the extent to which the required context information is free from errors and inconsistency. The external quality of a context refers the appropriateness of context information for a given application (i.e. its fitness-for-use).

In geospatial interoperability, context reasoning needs to verify and compare the degree of appropriateness of the contexts associated with different sources of information. Consequently, evaluating and modeling the external quality of context is important for context reasoning in geospatial interoperability.

In previous work we proposed a framework to overcome the conflicts related to the semantic interoperability of geospatial data. The framework is based on bidirectional communication and reasoning about context information [Sboui et al., 2007]. In another work, we proposed a set of indicators for the external quality of context and a method to evaluate those indicators [Sboui et al., 2009]. In this paper, we propose an ontology-based approach to model the external quality of a context with respect to the application for which

the interoperability is carried out. This approach aims at fine-tuning context reasoning and hence enhancing geospatial interoperability.

In the next section, we briefly present the existing approaches to reason about context. In Section 3, we present a set of indicators that have a major role in indicating the quality of context; and we propose a model to represent and reason about the external quality of context. We conclude and present further works in Section 4.

2 Approaches for context modeling and reasoning

Several approaches have been proposed for context reasoning (e.g., ontology-based reasoning, rule-based reasoning, and probabilistic reasoning). Ontology-based and rule-based reasoning are the two major approaches [Gu et al., 2004; Bikakis et al., 2007; Tang et al., 2007].

Ontology-based reasoning

Ontology, as a formal and explicit specification of a shared conceptualization, is considered as an efficient technique for modeling context enabling software agents to interpret and reason about context information [Gu et al., 2004; Souza et al., 2006; Frank, 2007]. Ontology-based approaches use Semantic Web technologies (e.g., RDF(S) and OWL) to model and reason about context information. These approaches are the most commonly used thanks to their formal structure and high expressiveness.

Rule-based reasoning

These approaches are based on predefined sets of rules that aim at verifying the consistency of context information [Bikakis et al., 2007]. They typically provide a formal model for context reasoning and can be integrated with the ontology-based reasoning approaches.

Both approaches focus on verifying the internal quality (e.g., consistency) of the context information. They pay less or no attention to the external quality of context information (i.e., fitness for use). However, in semantic interoperability, context reasoning needs to take into account the quality of context with regard to the application for which the

interoperability is carried out. Consequently, evaluating and modeling such quality is important to enhance context reasoning.

3 Evaluating and Modeling the external quality of production context

3.1. Evaluating the external quality of context

In previous work [Sboui *et al.* 2009], we proposed a restricted set of indicators and a method for evaluating the external quality of metadata with regard to a specific use. These indicators are: *convenience of language*, *completeness*, *trust*, and *freshness*. Each indicator is evaluated according to a function. The resulting quality value is within the interval [0, 1]. The value 1 indicates perfect quality while the value 0 indicates completely poor quality. Based on this value, a qualitative value (i.e., “good”, “medium” or “poor”) is assigned to the external quality of both explicit and implicit production context.

1- *Convenience of language*. It indicates the convenience of using a given language to represent the production context of geospatial data. For example, the convenience of a free natural language for a novice user is “medium”.

2- *Completeness*. It shows the quantity of the production context with regards to user’s requirements. We recognize thematic, spatial, and temporal completeness. For example, the spatial completeness of a context that does not contain information about reference system is “poor”.

3- *Trust*. It describes the degree of faith that we have about the production context transmitted in a chain of interveners in semantic interoperability of geospatial data. For example, the faith we have about data precision is “medium”.

4- *Freshness*. This quality indicator shows the degree of rationalism related to the use of context information at a given time. The value of freshness is determined by the age and

lifetime of the context information. For example, a context defined in 2008 is fresh (i.e., the freshness is “good”).

We should notice that these indicators do not aim at being complete or precise but rather at making agents globally aware of the external quality of the production context.

3.2. Modeling the quality of context

We propose a formal model based on ontology using Web Ontology Language (OWL) to represent the previously defined indicators of external quality and facilitate context reasoning. The model embeds both the production context and the information about its external quality using OWL classes and properties. The choice of ontology technique is motivated by the fact that ontology provides: (1) a flexible structure with explicit vocabulary to represent concepts and relations of both a context and its external quality, (2) logic reasoning mechanisms which are necessary to verify the context external quality, and (3) a common structure to exchange data between interveners in the geospatial interoperability.

Figure C.1 shows a simplified view of the model that includes a set of OWL classes such as *ContextElement*, *ExternalQuality*, *Indicator* and *Value*. *ContextElement* class allows representing any context element (spatial, thematic or temporal). *ExternalQuality* class allows specifying various qualities with different external quality indicators. *Indicator* class defines the indicators of the external quality of context in a specific use. *LanguageConvenience*, *Completeness*, *Trust* and *Freshness* are subclasses of *Indicator*. Each indicator has a *Value* that can be *Good*, *Medium* or *Poor*.

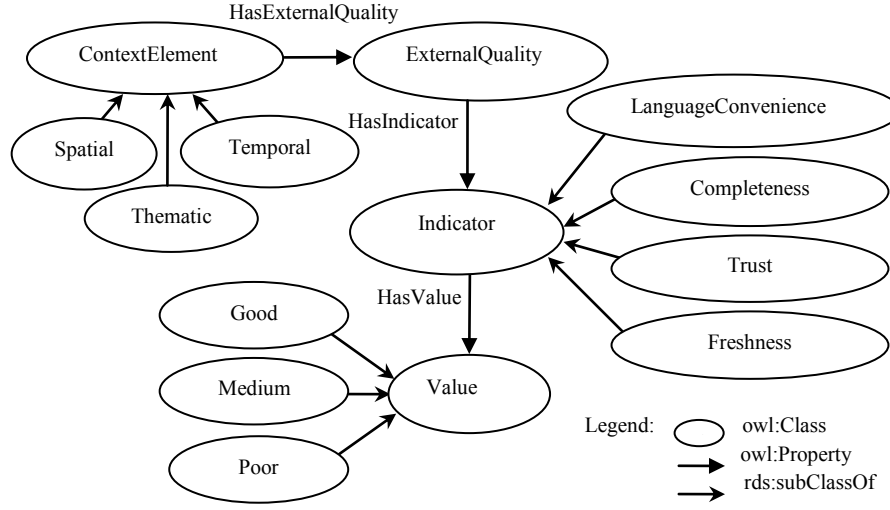


Figure C.1: OWL-based context external quality model

The semantics of OWL ontology are derived from Description Logics (DL). Description Logics are a family of formalisms for representing knowledge [Baader 2003]. DL consists of four parts: constructors which represent concepts and relations (or roles), terminological assertions, assertions about instances, and a set of rules which allow inferring new knowledge from existing one. We use the description logics *ALC* as an example. *ALC* supports Boolean constructors on concepts and roles, and universal and existential value restrictions on roles. Given a set of primitive concepts and a set of primitive roles, denoted with names from the set N_C and the set N_R respectively, additional concepts and roles can be constructed based on the following syntax rules:

$$C, D \rightarrow A | C \cup D | C \cap D | \neg C | \forall R.C | \exists R.C$$

In the proposed model, we defined the following primitive concepts and primitive roles:

Primitive concepts: $N_C = \{\text{ContextElement}, \text{ExternalQuality}, \text{Indicator}, \text{LanguageConvenience}, \text{Completeness}, \text{Trust}, \text{Freshness}, \text{Value}, \text{Good}, \text{Medium}, \text{Poor}, \text{SpatialContextElement}, \text{TemporalContextElement}, \text{ThematicContextElement}\}$

Primitive roles: $N_R = \{\text{HasName}, \text{HasQuality}, \text{HasInternalQuality}, \text{HasExternalQuality}, \text{HasIndicator}, \text{HasValue}\}$

Using primitive concepts and primitive roles, we can define additional concepts and roles. For example the concepts: `BadExternalQuality`, `GoodExternalQuality` can be defined as follows:

$$\text{BadExternalQuality} \equiv \text{ExternalQuality} \cap \forall \text{HasIndicator} (\text{Indicator} \cap \exists \text{HasValue.Poor})$$

That is, a bad external quality is an external quality that has poor value for all its indicators.

$$\text{GoodExternalQuality} \equiv \text{ExternalQuality} \cap \forall \text{HasIndicator} (\text{Indicator} \cap \exists \text{HasValue.Good})$$

Also, in order to facilitate the comparison of heterogeneous context elements, we define the following additional concepts: `ExcellentContextElement` and `BadContextElement`. These two concepts indicate, respectively, that a context element has a good and bad quality for all indicators. They can be represented as follows:

$$\text{ExcellentContextElement} \equiv \text{ContextElement} \cap \forall \text{HasExternalQuality.GoodExternalQuality}$$

$$\text{BadContextElement} \equiv \text{ContextElement} \cap \forall \text{HasExternalQuality.BadExternalQuality}$$

The above examples show that the proposed model allows inferring conclusions not only from existing context information, but also from information about the external quality of context. Based on such conclusions, interveners in the interoperability process (i.e., agent systems or humans) will be able to appropriately reason about the context of geospatial data and make appropriate decisions (e.g., comparing two heterogeneous context elements and considering one of them, or ignoring a context that has a poor external quality).

4 Conclusion

Modeling and reasoning about context information still remains a major issue in geospatial interoperability. Although many approaches have examined this issue, they focus solely on the internal quality of context. In this paper, we propose an approach based on ontology to model and reason about context taking into account the external quality of context. Such model aims at helping interveners in the geospatial interoperability to appropriately reason about context information. In addition, the proposed model provides relevant information

that can be used in making appropriate decisions. For example, if interveners have to choose between two heterogeneous context elements, they will be invited to choose the element with a better external quality (i.e. the element fitting better for the current use).

Further research is undergoing to define additional indicators of context external quality such as relevancy and granularity of context information. Then, a prototype will be implemented to validate the importance of the proposed model in enhancing the geospatial interoperability.

References

- [Baader, 2003] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P.F. Patel-Schneider. *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, London, 2003.
- [Bikakis *et al.* 2007] A. Bikakis, T. Patkos, G. Antoniou, and D. Plexousakis. A Survey of Semantics based Approaches for Context Reasoning in Ambient Intelligence. In *Proceedings of the Aml'07*, pages 15–24, 2007.
- [Bishr, 1998] Y. Bishr. Overcoming the semantic and other barriers to GIS interoperability. *Int. Journal of GIS*, 12(4):299–314, 1998
- [Brodeur, 2004] J. Brodeur. *Interopérabilité des données géospatiales: élaboration du concept de proximité géosémantique*. Ph.D. Dissertation. U. Laval, 2004.
- [Frank, 2007] A.U. Frank. Data Quality Ontology: An Ontology for Imperfect Knowledge. *COSIT 2007, LNCS 4736*, pages 406–420, 2007.
- [Gu *et al.* 2004] T. Gu, X.H. Wang, H.K. Pung, and D.Q. Zhang. An Ontology-based Context Model in Intelligent Environments. In *Proceedings of CNDS*, 2004.
- [Harvey *et al.*, 1999] Harvey, F., W. Kuhn, H. Pundt, Bishr Y. and C. Riedemann. Semantic Interoperability: A Central Issue for Sharing Geographic Information. *The Annals of Regional Science*, pp. 213-232, 1999.

- [Henricksen and Indulska 2004] Henricksen, K., and J. Indulska. Modelling and Using Imperfect Context Information. In: Proceedings of PERCOMW, 2004.
- [Sboui et al., 2007] Sboui T., Bédard Y., Brodeur J., and T. Badard. A Conceptual Framework to Support Semantic Interoperability of Geospatial Datacubes. ER Workshops 2007, LNCS 4802, pp. 378–387, 2007.
- [Sboui et al., 2009] Tarek Sboui, Mehrdad Salehi and Yvan Bédard. Towards a Quantitative Evaluation of Geospatial Metadata Quality in the Context of Semantic Interoperability. To appear in the proceedings of the 6th ISSDQ, 2009.
- [Souza et al., 2006] Souza, D., Salgado, A.C. and Tedesco, P.: Towards a Context Ontology for Geospatial Data Integration. Int. Workshop on Semantic-based Geographical Information Systems (SeBGIS'06), 2006.
- [Staub et al., 2008] Staub, P., H.R. Gnagi and A. Morf. Semantic Interoperability through the Definition of Conceptual Model Transformations. Transactions in GIS, 12(2):193–207, 2008.
- [Tang et al., 2007] Tang S., Yang J., and W. Zhaohu. A Context Quality Model for Ubiquitous Applications, 2007.


Annex D: Example of application of managing the risks of data misinterpretation based on the fitness-for-use of conceptual models

This Annex provides a complete version of the example introduced in section 5.5. This example shows how the risk management approach can be applied to fine-tune the process of semantic interoperability between geospatial datacubes. In this example, we continue with the two geospatial datacubes to be involved in the interoperability, shown of the Figure 3.8, and the requirements of end-users in the form of a geospatial datacube model, shown in Figure 5.2. 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 both datacubes.

We evaluate the indicators at each layer of the interoperability framework (cube, measure, dimension, hierarchy, level). To simplify the evaluation, we set the value 1 for a and b (which indicate the importance of each quality indicator of schema and production context respectively; see equations (6) and (7)), and that for all the levels. We should note that, in this example, we do not consider other elements of production context than those described in the metadata. Consequently production context is indeed metadata.


1. Cube layer:

When two datacubes source are semantically related (e.g., by their objective, their name or their contextualized measures, i.e., by taking into account their dimensions), then their schema and production context are analyzed, their quality indicators are evaluated and a decision is made about the continuation of the interoperability process.


In our example (c.f. Figure 3.8), the datacubes do not have a common subject of analysis. Accordingly, there is no semantic conflict between the geospatial datacubes. Then, the symbol  will be shown.

2. Measure layer:

When two measures are semantically related (e.g., by their name, their definition, their descriptive or geometric nature), then their structure (e.g., data type, value domain, geometric primitive) and their production context are analyzed, the quality indicators are evaluated and a decision is made about the continuation of the semantic interoperability.

In our example, the measures of both geospatial datacubes have not been defined for the same subject of analysis. Accordingly, there is no semantic conflict at the measure layer. The stakeholder is invited to continue to the next layer (dimension layer), then the symbol  will be shown.

3. Dimension layer:

If the dimensions of the two datacubes are not semantically related, the stakeholder does not have to go into more details for these dimensions, the symbol  will be shown and the stakeholder will be invited to continue with other possible combinations of dimensions. It is the case for dimensions *Age category* of datacube *C1* and *Region* of datacube *C2*.

If there is a semantic relation between two dimensions, such as the dimensions *Administrative region* of datacube *C1* and *Region* of datacube *C2*, then the quality of both dimensions is evaluated for their structure and their production context. This aims to facilitate making a decision about the choice of one dimension over the other.

a. Fitness-for-use of schema:

- *Relevance of the number of hierarchies (relevance of structure)*

The dimension *Administrative region* of the datacube *C1* has only one hierarchy: (*Country, Province, Territory* and *City*), whereas the dimension *Region* of the datacube *C2* has two hierarchies: (*Country, Province, Territory*, and *City*) and (*Country, State*, and *City*). We remind that end-users need only one hierarchy: (*Country, Province*, and *City*). Then, the quality indicator *Relevance of the number of hierarchies* is evaluated according to the formula (1):

$$\text{For } C1, P_s(\text{Administrative region, Administrative division}) = 1/1 = 1$$


For *C2*, since the number of hierarchies of dimension *Region* is larger than that of the semantically related dimension of the required datacube *C3* (*Administrative division*), then:

$$P_s(\text{Region, Administrative division}) = 1$$

According to the formula (6) the quality of datacubes schemas:

$$Q_s(\text{Administrative region, Administrative division}) = 1$$

$$Q_s(\text{Region, Administrative division}) = 1$$

Consequently, the symbol  will be shown for the two dimensions *Administrative region* and *region*. In the following layers (hierarchy and level), the analysis will be refined to lead to a choice among the dimensions of datacubes sources aiming to fit end-user requirements (expressed in model C3).

b. Fitness-for-use of production context:

➤ *Relevance of production context*

Production context associated with the dimension *Administrative region* of *C1* have two elements which are semantically related to the elements required by end-users (the spatial coverage and the year of creation of the dimension). On the other hand, production context associated with the dimension *Region* of *C2* have only one element which is semantically related to an element required by end-users (the spatial coverage of the dimension). If we consider that the level of importance of spatial information is 1, then based on formula (3):

$$P_m (\text{Administrative region}, \text{Administrative division}) = 2/2 = 1$$

$$P_m (\text{Region}, \text{Administrative division}) = 1/2 = 0.5$$

➤ *Freshness of production context*

Production context of the dimensions *Administrative region* of *C1* and *Region* of *C2* were created respectively in 2002 and 1982. Moreover, these two dimensions have the same lifetime: 30 years. Consequently, according to formula (4):




$$A_m (\text{Administrative region}, \text{Administrative division}) = 1 - (2005-2002/30) = 0.9$$

$$A_m (\text{Region}, \text{Administrative division}) = 1 - (2005-1982/30) = 0.23$$


According to the formula (7) the quality of production context:

$$Q_m (\text{Administrative region}, \text{Administrative division}) = (1 + 0.9) / 2 = 0.95$$

$$Q_m (\text{Region}, \text{Administrative division}) = (0.5 + 0.23) / 2 = 0.36$$

Consequently, the symbol  will be shown for the dimension *Administrative region*. On the other hand, the symbol  will be shown for the dimension *Region*. The stakeholder is then invited to be careful when considering the dimension *Region* in the process of interoperability. Moreover, the stakeholder is invited to evaluate the quality of the detailed levels of these two dimensions (i.e., hierarchy and levels) by taking into account the difference in quality of these dimensions. We should note that, if the quality of one of two dimensions was very poor, then the symbol  would have been shown and the stakeholder would have been advised not to consider the detailed levels of this dimension.

4. Hierarchy layer:

If the hierarchies of two dimensions are not semantically related, the stakeholder will not have to drill down into more details of these hierarchies (levels), the symbol  will be shown, and the stakeholder will be invited to continue with other possible combinations of hierarchies of two dimensions. On the other hand, if there is a semantic relation between two hierarchies, such as the hierarchies (*H1*: *City*,

Province, Territory and Country) of dimension *Administrative region* of datacube *C1* and (*H2: City, Province, Territory and Country*) of the dimension *Region* of the datacube *C2*, then we evaluate the quality of schema and production context of the two hierarchies.

a. Fitness-for-use of schema:

➤ *Relevance of the number of levels (relevance of structure)*

Each hierarchy (*H1* and *H2*) contains 4 levels. Since the hierarchy (*H3: City, Province, Territory and Country*) of the dimension *Administrative division* of *C3* contains 4 levels, according to the formula (1):

$$P_S(H1, H3) = P_S(H2, H3) = 4/4 = 1$$

➤ *Relevance of the order of levels (relevance of structure)*

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 datacube C3* has the following order: (*City < Province*) and (*Province < Country*).

For the datacube *C1*, according to expressions 2a and 2b (c.f., section 5.3.2.1):

The levels *City, Province* (or *Territory*) and *Country* (*C1*) are respectively, semantically related to the levels *City, Province* and *Country* (*C3*). Moreover, the level orders *City < Province* (in *C1*) and *City < Province* (in *C3*), 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:


$$O_S(H1, H3) = (1+1+1+1) / 4 = 1$$

Similarly, the relevance of the hierarchy (*City, Province, Territory and Country*) of the datacube *C2* is calculated based on expressions (2a and 2b):

$$O_S(H2, H3) = (1+1+1+1)/4 = 1$$

Finally, according to the formula (6):

$$Q_S(H1, H3) = Q_S(H2, H3) = 1$$

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

b. Fitness-for-use of production context:

➤ *Relevance of the production context*

Production context associated with the hierarchy (*City, Province, Territory* and *Country*) of the datacube *C1* contain an element which is semantically related to the element required by end-users (using the “multiple representation” for data insertion in geospatial levels). On the other hand, the production context associated with the hierarchy (*City, Province, Territory* and *Country*) of *C2* does not contain any element that is semantically related to the element required by end-users. If we consider that the level of importance of spatial information is 1, then, according to the formula (3):



$$P_m(H1, H3) = 1/1 = 1$$

$$P_m(H2, H3) = 0/1 = 0$$

Also, according to the formula (7):

$$Q_m(H1, H3) = 1/1 = 1$$

$$Q_m(H2, H3) = 0/1 = 0$$

Therefore, the symbol  will be shown for the hierarchy (*City, Province, Territory* and *Country*) of *C1*. The symbol  will be shown for the hierarchy (*City, Province, Territory*, and *Country*) of *C2*. Therefore, the consideration of the second hierarchy risks harming the interoperability between geospatial datacubes.

Thus, the stakeholder is invited to continue to evaluate only the levels of the hierarchy (*City*, *Province*, *Territory* and *Country*) of *CI*.

5. Level layer:

a. Fitness-for-use of schema (Level City of CI):

➤ *Relevance of the number of attributes (relevance of structure)*

The number of attributes of the level *City* of the datacube $CI = 1$ (*City-name*). Since end-users need 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 *CI*, each member of the level *City* is represented by a point (0D), whereas end-users need a polygon (2D) to represent the members of the same level *City*. According to the table 5.1:

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

Then, according to the formula (6), the structure quality of the level *City* of *CI*:

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

Consequently, the quality is reasonably satisfactory, then the symbol  is shown to the stakeholder to inform him/her about potential risks that may occur when considering this level. Based on the proposed framework, the stakeholder can decide to solve the problems related to this level or to endure the potential consequences of these problems.


b. Fitness-for-use of production context (Level City of CI):

➤ *Relevance of the production context*

Production context of the datacube *CI* contain two elements which are semantically related to the elements required by end-users (the spatial referencing system and the scale of representation). If we consider that the importance of spatial information is 1, then according to the formula (3):

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

According to the formula (7), the quality of the production context of the level *City* of *CI*: $Q_m(City, City) = (1+1)/2 = 1$

Consequently, the structure quality of the level *City* of *CI* is quite satisfactory. The symbol  will be shown to the stakeholder.

Similarly, the qualities of the levels *Province* and *Country* of the datacube *CI* are evaluated based on the formulas (6) and (7):

$$Q_s(Province, Province) = Q_m(Province, Province) = 1$$

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

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. It is the case of dimensions *Age category* and *Forest stand* in our example.

Figure D.1 shows an example of a common model that could be obtained to enable the interoperability between geospatial datacubes *CI* and *C2* according to the different levels of the general framework (see Figure 5.3) and using the proposed quality indicators.

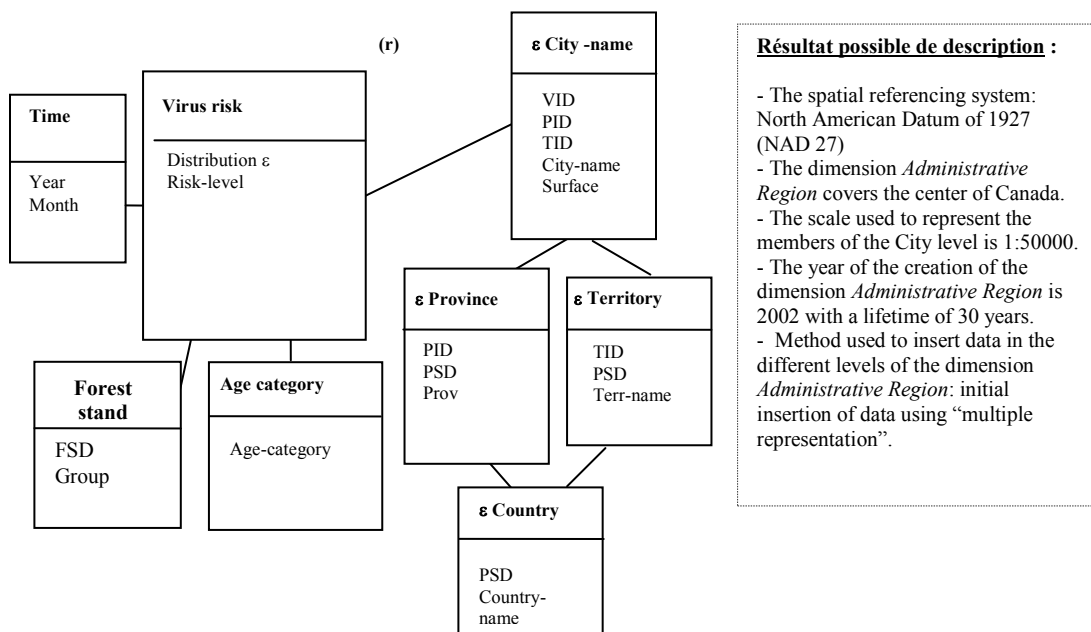


Figure D.1: Example of final common model resulted from the interoperability between geospatial datacubes interoperability.

The common model is obtained after dealing with the risk of data misinterpretation and, hence, can considerably facilitate the navigation task. Such model is useful especially in emergency situations (i.e., forest fire that affects adjacent countries) when end-users need to rapidly navigate through data stored in the geospatial datacubes without preoccupying themselves with the problems of heterogeneity. In such situation, and using the resulting model, end-users are able to navigate different geospatial datacubes, developed in these countries, in order to get the right information and act quickly to avoid catastrophic consequences.