uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Juan Xia He
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Geography)
GRADE / DEGREE

Department of Geography
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

An Ontology-Based Methodology for Geospatial Data Integration

TITRE DE LA THÈSE / TITLE OF THESIS

M. Sawada
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

H. Cao

K. Gajewski

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

AN ONTOLOGY-BASED METHODOLOGY FOR GEOSPATIAL DATA INTEGRATION

By

JUAN XIA HE

A thesis submitted to
The Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of
Master of Science, Geography

Department of Geography
UNIVERSITY OF OTTAWA

Your file Votre référence
ISBN: 978-0-494-73790-3
Our file Notre référence
ISBN: 978-0-494-73790-3

ABSTRACT

Data semantic and schematic heterogeneity is a major obstacle to the reuse and sharing of geospatial data. This research focuses on developing an ontology-based methodology to logically integrate heterogeneous geographic data in a cross-border context. Three main obstacles hindering data integration are semantic, schematic, and syntactic heterogeneity. Approaches to overcome these obstacles in previous research are reviewed. Among the different approaches, an ontology-based approach is selected for horizontal geospatial data integration in the context of cross-border applications. The integration methodology includes the extraction of application schemas and application ontologies, ontology integration, the creation of a reference model (or ontologies), schema matching and integration, and the creation of usable integrated datasets. The methodology is conceptual and integrates geospatial data based on the semantic content and so is not tied to specific data formats, geometric representations, or feature locations. In order to facilitate the integration procedure, four semantic relationships are used: refer-to, semantic equivalence, semantic generalization, and semantic aggregation. A hybrid ontology approach is employed in order to facilitate the addition of new geospatial data sources to the integration process. As such, three levels of ontologies are developed and illustrated within a MS ACCESS database: application, domain, and a reference model. Furthermore, a working integration prototype is designed to facilitate the integration of geospatial data in the North American context given the semantic and schema heterogeneities in international Canadian-US geospatial datasets. The methodology and prototype provide users with the ability to freely query and retrieve data without knowledge of the heterogeneous data ontologies and schemas. This is illustrated via a case study identifying critical infrastructure around the Ambassador Bridge international border crossing. The methodology and prototype are compared and evaluated with other GDI approaches and by criteria introduced by Buccella et al. (2009). Specific challenges unique to GDI were uncovered and include geographic discrepancies, scale compatibility and temporal issues.

ACKNOWLEDGEMENTS

This thesis would not have been possible without help and support from many people, not all of whom I can mention here.

I gratefully acknowledge the love and support provided by my husband Jian and my son Yiyang. Thanks for their encouragement and confidence in my ability to finish the thesis.

I also gratefully acknowledge my supervisor, Dr. Mike Sawada, for the encouragement, advice, support, and guidance that you have given me over the years. I gratefully acknowledge the guidance of my committee members Dr. Huhua Cao and Dr. Konrad Gajewski for their guidance and support.

I also thank my parents and in-laws for all the love and support they have given me throughout my graduate studies.

I wish to acknowledge all my friends and fellow students with whom I have had the pleasure of working with in the Laboratory for Applied Geomatics and GIS Science (LAGGISS). Thanks you everyone for all the great advice and encouragement regarding my thesis. A special thanks to Ruibo Han and his wife LinNa, Jing Feng, Claudia Fall, Zoran Reljic, Melanie Zahab, and Marian Kremer.

Last, but not least, I would like to acknowledge the academic, financial, and technical support provided by the University of Ottawa, and in particular the Department of Geography.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ADO -- Active Data Objects

CGDI -- Canadian Geospatial Data Infrastructures

CGVD28 – Canadian Geodetic Vertical Datum of 1928

CIP -- Critical Infrastructure Protection

DBMS -- Database Management System

DHS GDM -- Department of Homeland Security Geospatial Data Model

DOI -- Data of Interest

EPCIP -- The European Programme for Critical Infrastructure Protection

EPOI -- Enhanced Points of Interest

FGDC -- Federal Geographic Data Committee

GCS – Geographic Coordinate System

GDI -- Geospatial data integration

GIC -- Geographic Information Community

GIS -- Geographic Information System

GML – Geography Markup Language

GSDI -- Global Spatial Data Infrastructure

GUI -- Graphical User Interface

ISO -- International Organization for Standardization

NAD83 -- The North American Datum of 1983

NAD83CSRS – The North American Datum of 1983 in Canadian Spatial Reference System

NAVD88 – The North American Vertical Datum of 1988

NSDI -- National Spatial Data Infrastructure

NSIF -- National Spatial Information Frameworks

ODBC -- Open Database Connectivity

ODGIS -- Ontology-driven GIS

OGC -- Open Geospatial Consortium

OIL -- Ontology Inference Layer

RDF -- Resource Description Framework

RS – Remote Sensing

SLD -- Styled Layer Descriptor

SQL -- Structured Query Language

SVG -- Scalable Vector Graphics

TOC – Table of Content of ArcMap

UML -- Unified Model Language

VBA -- Visual Basic for Application

WCS -- Web Coverage Service

WCT -- Web Coordinate Transformation Service

WFS -- Web Feature Service

WMF -- Web Map Service

WOI -- Where of Interest

XML -- Extensible Markup Language

CHAPTER 1

Introduction

Geographic (or location-related, geo-referenced) data is increasingly involved in governmental administration, public services, and commercial activities, and today individual daily lives are affected via the proliferation of internet mapping and global positioning technology. In addition, in many fields of research, there is a strong and significant demand to share and use existing geospatial data available from different national and sub-national authorities within seamless GIS applications. However, because such datasets are produced and maintained independently and autonomously by numerous organizations or data authorities for various purposes, geospatial data that are conceptually meant to represent the same 'real world' entities (e.g., roads) are prone to be heterogeneous and incompatible in many aspects. Specifically, such data heterogeneity can be conceptualized from syntactic, schematic, and semantic viewpoints and these aspects of heterogeneity result in significant issues when trying to utilize datasets in seamless GIS applications, especially in cross-border contexts. When reusing various existing geospatial datasets that cross administrative boundaries, in particular, international ones, users have to tackle issues of data heterogeneity before reassembling the datasets.

1.1 Geospatial Data Interoperability

As an approach for sharing geospatial data and reducing costs, geospatial data interoperability has been a significant research topic in geographic information communities (GICs) since the beginning of 1990s (Abel et al., 1998; Bishr, 1998; Brodeur et al., 2003; Fonseca et al., 2003; Fonseca et al., 2002). Data interoperability provides the ability to access and use heterogeneous data among different systems and requires knowledge or agreement on application schema concepts (Bishr, 1998; Devogele et al., 1998; FILETO, 2001; Stoimenov and Djordjevic-Kajan, 2005). Shown in Figure 1.1, data interoperability has a technical level and a semantic (or information) level of interoperability from a system and data model perspective, respectively (Bishr, 1998; Stoimenov and Djordjevic-Kajan, 2005; Uitermark, 2001). Technical level interoperability is related to information accessing issues, such as network protocols and standards for datasets, which make data access from remote systems possible; semantic level interoperability is related to the understanding of information semantics which enables the accessed data to be processed and interpreted (Bishr, 1998; Uitermark, 2001; Stoimenov and Djordjevic-Kajan, 2005).

Aside from the technical and semantic perspectives, Parent et al. (2000, p222) suggested that data interoperability can be classified in three ways. At the lowest or basic level there is no integration and this level enables one DBMS (Database Management System) to request and obtain data from another one. A second approach or intermediate level of integration supports user-driven access and/or integration of data from multiple databases where users need to manipulate the local datasets consistently via a multi-database query language or by defining their external schemas. Finally, the third or highest level involves the provision of integrated access to diverse datasets without data duplication, formatting or transformation. Other researchers like Bishr (1998) have introduced six levels of interoperability between two or more independent GISs (Geographic Information Systems) but these additional levels are encompassed by Parent et al.'s (2000, p222) integration levels described above (shown in Figure 1.1). However, one significant difference exists: Bishr (1998) adds to the highest level (sixth level) of integration ability to provide a seamless communication between independent GISs without having prior knowledge of their underlying semantics. Bishr's (1998) modification of Parent et al.'s (2000, p222) highest level of integration represents the ultimate goal of geospatial data interoperability. In that sense, at that highest level, two cross-boundary geospatial datasets become a single entity for query and application based analyses.

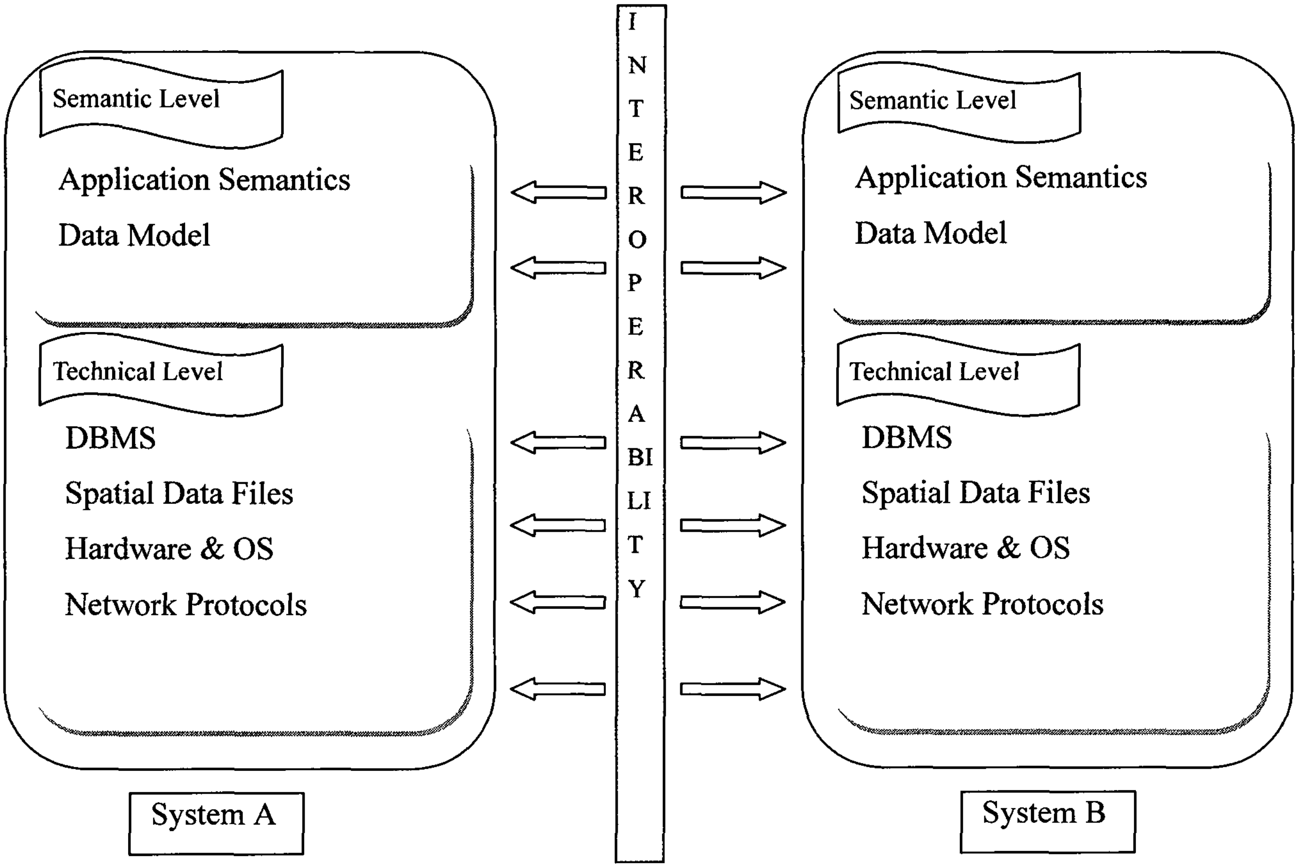

Figure 1.1: The technical level and semantic Level of data interoperability (adapted from Bishr, 1998).

1.2 Geospatial Data Integration

Geospatial data integration (GDI), at the semantic level, is the key to geospatial data interoperability (Parent et al., 2000, p221; Uitermark, 2001). By defining a virtual database, Parent et al. (2000) considered data integration as a logical unification process for distributed and heterogeneous datasets. To achieve data integration, two fundamental aspects, schema mapping generation and query responses, need to be addressed (Souza et al., 2006). Moreover, for convenient use, geospatial data integration needs to provide users with a uniform graphic interface to access and retrieve information from diverse data sources (Souza et al., 2006). Data integration also has five levels that deal with hardware, platform, syntactics, semantics (Izza, 2009), and schemas. This research will focus on semantic integration in order to achieve schema integration.

Geospatial data integration involves unique issues that are not found when integrating non-geospatial data. First, the data to be integrated represent specific geometry types such as points, lines, or polygons. Therefore, when presenting the retrieved results (such as shapefiles) to users, semantically similar objects residing in various datasets need to be grouped by geometry type. Next, geospatial data integration must address scale and/or resolution compatibility issues since both the representation of a real world observation as a particular geometry type and the spatial analysis models are both scale dependent. Finally, in addition to a hierarchy relationship such as generalization and aggregation, more space-based relationships are involved in geospatial data integration, that is to say, topology is paramount. For example, road nodes are connected to road segments; one golf course is adjacent to a school; a landfill site is near a highway, etc. Therefore, geospatial data integration requires more attention on geometry types, data resolutions, and topological relationships that are accounted for in the data topology considerations with non-geospatial data.

The integrated geospatial data can be applied in numerous contexts, including geospatial data update propagation (Uitermark, 2001; Wang & Wei, 2008), within application areas that need spatially coincident or mutually exclusive geospatial data, for example, within the field of border security planning and monitoring (The Carbon Project, 2009), public security and safety (Gugliotta, 2006), public health (Moreno-Sanchez et al., 2007; Gao et al.,2008), environment (Allen et al., 2008; Pundt & Bishr, 2002; Lin & Ludäscher, 2006), and resource management (Allen et al., 2007). In cross-boundary applications like border situations, integration is a necessary concern. Borders clearly determine geographic boundaries of political entities or legal jurisdictions, such as borders of a nation, a province, a state, a city, or a town, and many spatial phenomena like man-made disasters, environmental pollution, and disease diffusion that ignore the existence of political or administrative borders. In such situations, consistent and integrated geospatial data that cross borders is a powerful tool to efficiently address such issues.

1.3 Cross-border Geographic Dataset Integration

There are many reasons for two countries that share a border to share their geospatial data along the bordering region. The USA and Canada are two countries that share 8891 kilometres of international border, the longest undefended border in the world (International Boundary Commission, 2008). Canada and the USA have created their respective geospatial data infrastructures (i.e., Canadian Geospatial Data Infrastructures (CGDI) and U.S. National Spatial Data Infrastructure (NSDI)) based on the Open Geospatial Consortium (OGC) Standards to enable users develop and exchange geospatial data within their respective countries but to a lesser extent across their boarders (GeoConnections, 2006). Thus, users generally need not stop at the Canada-US border when exchanging geospatial data. However, in order to better coordinate border security programs, manage cross-border natural resources, respond to disasters, protect the environment, and streamline trade, users may want to freely access geospatial data along the border as well as process and understand the meaning of the data in order to combine and use these data. Some users may even ask for a unified data access interface to retrieve the geospatial data from the two countries. In short, cross-border geospatial data usage requires the data from the bordering countries be accessed and interpreted as an integrated set without digging into the detail of the 'other' country's data knowledge such as how they are stored and what the other's meanings may entail.

GDI in the context of cross-border geographic space takes two directions (Figure 1.2): one direction integrates geospatial data over the same geographical space and is named *Vertical Integration*; for example, the integration of Ottawa road data at different base scales (such as 1:10,000 and 1:25, 000 or the same topic datasets from different collection dates or authorities). Uitermark (2001) also presented an example of vertically integrating GBKN and TOP10 Vector data which were two topographic datasets; another direction is to integrate geospatial data crossing borders (such as trans-national borders, trans-provincial borders) named *Horizontal Integration*; for example, the integration of Canadian national road data and USA national road data. In the process of horizontal integration, geospatial data to be integrated are very prone to be heterogeneous in the facets of semantics, schema, and syntactics induced by differences of political, regulatory, standards, as well as variations in culture and technology development. By way of illustration, Figure 1.3 shows one dataset named "State" for the USA and another dataset named "Province" for Canada. Dataset State uses the concepts of ***state_name***, ***state_Fips***, ***sub_region***, ***state_ABBR***, etc. to represent the information of each state in the USA; however, dataset Province uses the concept ***Name***, ***Type_E***, and ***SRC_Agency*** to depict every province in Canada. These two datasets depict the information for the most compatible political units yet have heterogeneous schemas and semantics. In this horizontal situation, to attain integration one must overcome the heterogeneities and provide users with a seamless query interface such as a Graphical User Interface (GUI) to retrieve the required data without knowing the different semantics and schematics in the two datasets.

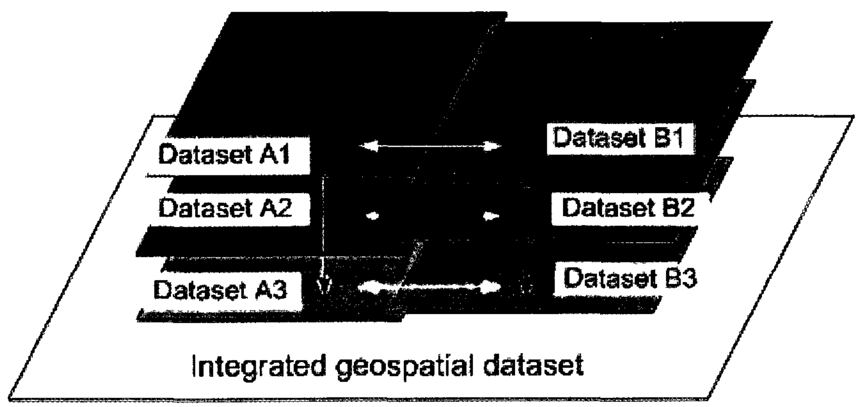

Figure 1.2: Integrating dataset A1, A2 and A3 or dataset B1, B2, and B3 is vertical integration, and integrating dataset A1 and B1, A2 and B2, A3 and B3 is horizontal integration.

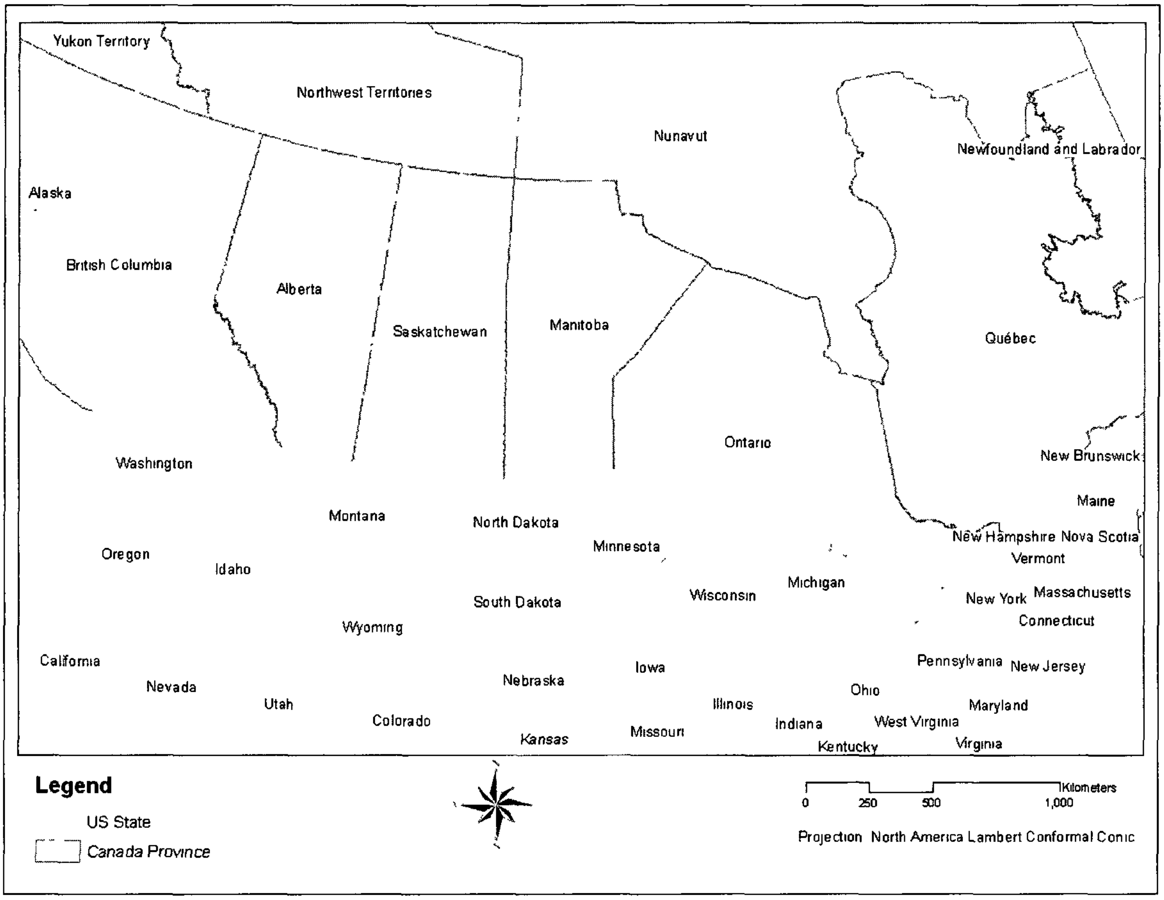

Figure 1.3: An example of geospatial datasets with heterogeneous schemas and semantics along the borders of Canada and the USA.

1.4 Research Objective

The objective of the thesis is first to develop an ontology-based methodology to overcome semantic heterogeneity for horizontal integration of geospatial data in the context of continuous geographic regions along trans-national borders and second to produce a case study for the integration of geospatial data crossing the continuous area along the borders of Canada and the USA. This objective will be addressed via the following research questions:

1) Can integration be undertaken effectively in the North American context given the semantic heterogeneity and schema heterogeneity presented in two diverse

international datasets? If yes then the outcome of this research will be an ontology-based geospatial data integration methodology and integrated framework datasets.

2) Can integration be streamlined via a software solution? That is to say, can a graphical user interface be developed that allows users to seamlessly pose queries and retrieve the information from the logically integrated datasets?

1.5 Thesis structure

To support the relevance of these questions and to propose a methodology, this thesis is organized as follows.

Chapter 2 is literature review that examines the three facets of geospatial data heterogeneity (i.e., semantic, schematic, and syntactic heterogeneity) and proposed solutions that have been successful in dataset integration.

Chapter 3 provides the definition of ontology and then presents an ontology-based geospatial data integration approach introduced in the literature. This integration approach includes two main steps: ontology exploration and schema mapping.

Chapter 4 introduces the methodology designed for horizontal geospatial data integration in the context of cross-border areas. This chapter first introduces the integration process, and then explains the detail of each step in the integration process which includes the development of application schemas and application ontologies, the construction of the domain ontologies and the reference model and schema integration. Four kinds of semantic relationships are introduced in the chapter: refer-to, semantic equivalence, semantic generalization, and semantic aggregation. The last part of this chapter presents the conceptual framework of the ontology-based geographic information integration.

Chapter 5 explains the content of framework datasets in the research, and then presents the procedures of the application of the integration methodology from Chapter 4 to integrate the framework datasets along the Canada-US border regions. This chapter also explains how to use MS ACCESS DBMS to record the ontologies of the concepts and relationships residing in the application schemas and the reference model. The last part of this chapter presents the development of a GUI seamless query and retrieval prototype.

Chapter 6 presents a case study: the application of the designed methodology and the prototype to identify the critical infrastructure around the Canada-US border areas. It starts with the importance and content of critical infrastructure data; then presents how the framework datasets and the prototype can be used to carry out the identification of critical infrastructure along the Canada-US bordering areas.

Chapter 7 discusses the advantages and limitations of the methodology and prototype application. Comparisons among different GDI approaches are made and suggests the use of standardized web services to upgrade the designed stand-alone prototype to an internet-based geospatial data integration service. The methodology and prototype are then evaluated by the criteria introduced by Buccella et al. (2009). The last part of this chapter lists some challenges that occurred in the research: geographic discrepancies, scale compatibility issue, and temporal discrepancies.

Chapter 8 presents conclusions and needed future research.

CHAPTER 2

Syntactical, Schema, and Semantic Heterogeneity in Geographic Data Integration

A geographic information community (GIC) is a group of users who share geographic information language and spatial feature definitions and communicate via a common set of basic concepts (Stoimenov and Djordjevic-Kajan, 2005). People in the GICs recognize real world facts and give them names and categorize them and then present them in a database using a specific model, e.g., schemata, complied with the database's underlying syntax, which is depicted in Figure 2.1 (OGC, 2009). Therefore, in the process of abstracting real world observations into a GIS domain, different GICs with various purposes may result in different schemes of the real world, known as ***schematic heterogeneity***. ***Semantic heterogeneity*** involves a single real world observation or entity having several descriptions or representations, such as different names, to meet different requirements. ***Syntactic heterogeneity*** arises when modeling such facts into a database: the database may be implemented in a different DBMS of different paradigms, for example, in a relational DBMS or in an object-oriented DBMS. When people reassemble geospatial data produced or maintained by different organizations or created for different purposes, the three facets of heterogeneity: syntactic, schematic, and semantic, are significant barriers for integrating geospatial data (Brodeur et al., 2003; Abel et al., 1998). This section will identify the attribute of the heterogeneities in semantic, schematic and syntactic aspects in more detail.

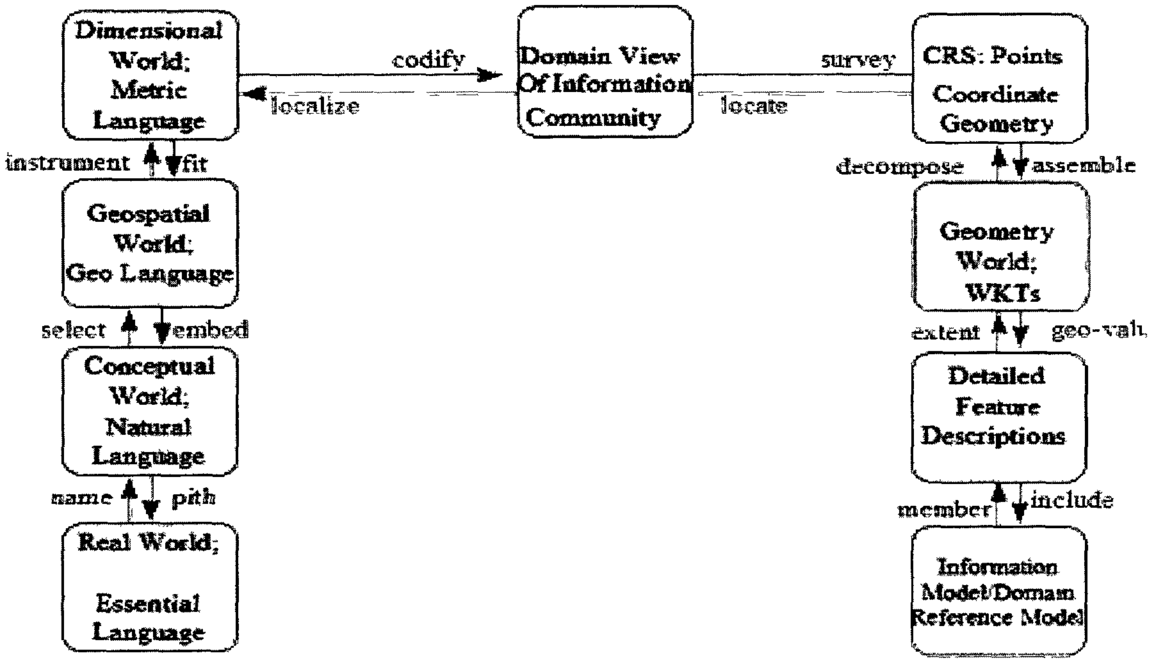

Figure 2.1: The process how people model the reality into GIS domain using schematics (OGC, 2009)

2.1 Semantic Heterogeneity

Semantics refers to the relationship among the representations in datasets and the corresponding features in reality within a certain context (Bishr, 1998). This definition is similar to the meaning of semantics in Linguistics which is the relationship between signs and symbols and what they represent. Semantic heterogeneity is induced when a real world fact/entity/observation has more than one representation or description in two or more datasets that need to be integrated. Semantic heterogeneity arises due to the different application purposes or perspectives embodied in the datasets from different sources (Bishr, 1998). Goh (1997) suggests three main causes for semantic heterogeneity (Goh's study (as cited in Wache et al., 2001)):

1) confounding conflicts, which is result from confounding of concepts, for example data items seem to have the same meaning, but actually differ in reality;
2) scaling and unit conflicts that refer to use different units of measures or scales of representation (generalization), for example using different reference systems when measuring a value or representing entities at different map scales resulting in more or less relative generalization among datasets; and,
3) naming conflicts, which include synonyms and homonyms, for example different naming schemas for information for the same observed entity.

Semantic heterogeneity often occurs at data/instance level and schema level (Souza et al., 2006). Different perspectives or conceptualizations of the real world will cause these two levels of semantic heterogeneity.

Data level semantic heterogeneity occurs when different geospatial datasets give the same names to different phenomena (like homonyms), or different names to the same phenomena (like synonyms). For example, in Canada road network data downloaded from www.geobase.ca "***Ramp***" is defined as inter-connecting roadways providing for the controlled movement between two or more roadways and includes not only highways, but also other kinds of roads such as main roads. While in the USA TIGER/Line data "***Ramp***" is defined as a road that allows controlled access from adjacent roads onto a limited access highway. Thus, the value "***Ramp***" appears in these two datasets but has different meanings in each and *prima-facia* can lead to integration errors.

Different representations for the same factual attributes in the real word will cause schema level semantic heterogeneity. Taking roads as an illustration: for the information of a crossroad (Figure 2.2), road management sections may focus on the pavement condition of the road (such as representation 1). While, transportation sections may be interested in the network of the roads (such as representation 3). In a small-scale topographic map, the crossroad may be depicted as representation 2. These different conceptualizations of the crossroad will cause different descriptions for the same crossroad in reality which will result in different representations and thus a heterogeneous data model. When such data are combined, these different representations (that include different data models, data structure, names, and so on) do refer to the same real-world or observable object. Therefore, integrating geospatial data firstly needs to solve semantic heterogeneity which has proved to be a significant and difficult barrier to reconcile (Bishr, 1998, Souza et al., 2006), especially for a cross-border geospatial applications (Allen et al., 2008). The goal of semantic integration is to create an explicit formal and shared vocabulary in the GIC which can form the basis of generating schema mapping.

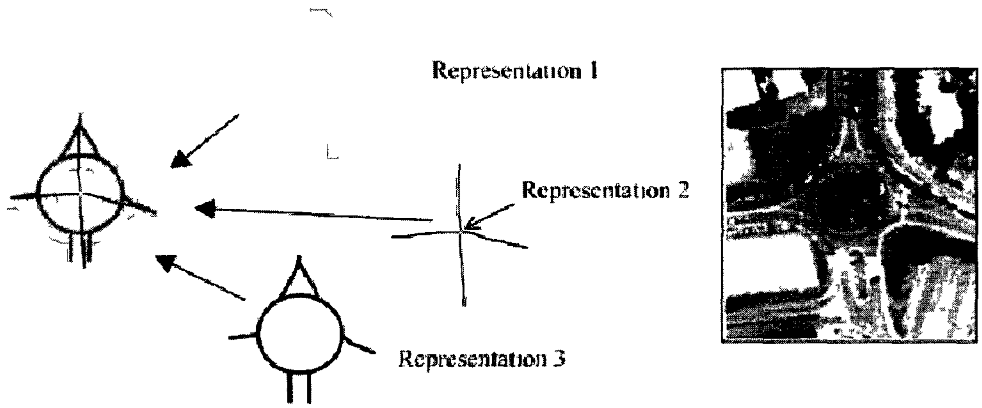

Figure 2.2: Alternative representations of the same crossroad (Balley et al., 2004).

2.2 Schematic Heterogeneity

Schemas of datasets are the definitions of logical structures or patterns that convey data (Hakimpour and Timpf, 2001), which include the definitions of classes, attributes, relationships among the classes, constraints for classes and their attributes. Schematic heterogeneity, also named structural heterogeneity (Brodeur et al., 2003), refers to different structures of data models, which may arise from differences in classification, class relationships, or class naming conventions. Bishr (1998) considers schematic heterogeneity as occurring when an object class in one database is considered as a property in another database, or when object classes have different aggregation or generalization hierarchies even when they describe the same real world entities. For example, Canadian roads are depicted by road segments, ferry connection segments, blocked passages, toll points, and junctions in a total of five feature classes, whereas the

USA roads are represented in one feature class: road segment, where a ferry connection segment is one kind of road segments (subtype), which is a property or attribute of a road segment.

Goh (1997) suggests that schematic heterogeneity comes from data type conflicts, labelling conflicts, aggregation conflicts, and generalization conflicts. Data type conflicts refer to the use of different primitive data or geospatial (geometric) types to represent data values (Goh, 1997). Take the example shown in Figure 2.2, a crossroad may be stored in road management datasets as a polygon, while in the transportation management datasets the same crossroad is recorded as a polyline, which is a kind of spatial data type conflict and is specifically one of data model representations that can be scale induced. Labelling conflicts refer to synonyms and homonyms when occurring at the schema element level (Goh, 1997), like synonymous class names in different datasets. Aggregation conflicts occur, for example, when one object in a dataset corresponds to a set of objects in the other dataset (Goh, 1997; Devogele et at., 1998). In GICs different mapping scales require a different aggregation level, which is a source of aggregation conflict. The different representations for the crossroad in Figure 2.2 is a good example of an aggregation conflict: transportation management may represent a crossroad with the lanes of the road, a polygon circle, and ramps, while in a smaller-scale topographic dataset, the same crossroad is aggregated into simple line segments or polylines with an intersecting point or junction that represents the geographic position of the centre-lines of the roads and the center point of the circle. Generalization conflicts differ from aggregation conflicts in that designers make different choices to determine how different classes should be encoded on a map (Goh, 1997). For example, all the road segments including ferry roads are stored in one feature class in the USA road datasets, while Canada road datasets have a road segment feature class and ferry road segment feature class, and these two different feature classes have identical definitions of class attributes. These conflicts of data type, labelling, aggregation, and generalization cause the schematic heterogeneity.

In order to solve schematic heterogeneity, one first needs to solve semantic heterogeneity (Abel et al., 1998). As described above, semantic heterogeneity may be related to schematic level heterogeneity. For example, dataset A and B are all describing water information in an area. In dataset A class ***River*** contains the information of streams, while, the same information is represented in class ***Water*** in dataset B. In order to share these two datasets, one needs to explore the relationship between class ***River*** and class ***water*** based on what they refer to in reality. Therefore, schematic integration can be achieved by means of schematic-level semantic integration. The target of schema integration or schema synthesis is to generate schema mapping via, for example, building up a single logically unified schema (for example a global view of schema) (Abel et al., 1998; Bishr, 1998; Rahm & Bernstein, 2001) which can be the basis of diverse data access and retrieval.

2.3 Syntactic Heterogeneity

Syntactic heterogeneity stems from using different kinds of data models to represent database elements and their underlying DBMS, for example, using a relational vs. object-oriented data model, or such heterogeneity can be caused by the use of different geometric representations of geographic objects in the datasets, for example, raster vs. vector representations (Bishr, 1998; Lutz and Klien, 2006). Syntactic heterogeneity is also observed in the progresses of, for example, GIS format translation, geospatial data structure transformation, and geometric data-type definition (Brodeur et al., 2003). As such, one can loosely think of syntactic heterogeneity as a 'data format' issue, an issue which is well addressed in today's software environments.

2.4 Solutions to Geospatial Data Heterogeneity

Syntactic and some aspects of schema heterogeneity, such as the data-type definition, different data models, and systems, can be overcome by universal standards and specifications (Lutz and Klien, 2006). The universally endorsed standards and specifications in the geospatial domain are also the base of geospatial information reuse and sharing. The main standardization bodies: the ISO (the International Organization for Standardization) Technical Committee 211 (ISO/TC 211) and the OGC are trying to overcome data model and system heterogeneity to accomplish geospatial data interoperability and integration (Lutz and Klien, 2006). The ISO develops *de jure* standards, whereas the OGC drafts *de facto* standards by a consensus process (Klopfer, 2005). The OGC is a non-profit, international industry consortium aiming to develop publicly available interface specifications for geospatial information and services interoperability (OGC, 2008). In addition to the ISO and the OGC, many organizations and countries, for example the European Union, Geoconnections of Canada, and the Federal Geographic Data Committee (FGDC) of the USA, endorse the standards and specifications prescribed by the ISO and the OGC, and then build up their own geospatial data infrastructures (for example the Canadian Geospatial Data Infrastructure (CGDI)). In practice, these standards and specifications are the foundation of geospatial data integration. For example, to tackle syntactic and schema heterogeneities, the GeoNis system developed by Stoimenov and Djordjevic-Kajan (2005) used a tool to transfer one data model of a local information source into a common data model provided by OGC standards.

However, even if all datasets and systems followed the aforementioned universal standards and specifications to represent the real world, semantic and schema heterogeneity are still significant problems facing geospatial data integration. For example, users can revise the standard data model to create an application schema to meet specific or ad-hoc requirements. Some approaches are developed in the literature to overcome semantic and schema heterogeneity to accomplish GDI. For example, the development of the Geospatial Semantic Web by the OGC (OGC, 2006), the federated data model approach (Sheth & Larson, 1990; Devogele et al. 1998), the Rough and Fuzzy geographical data integration approach (Ahlqvist et al., 2003), the statistical approach (Walter & Fritsch, 1999), the feature-based approach (Samal et al., 2004), the location-based approach (Safra et al., 2010), the object-oriented modeling approach (Gärtner et al., 2001; Abel et al., 1998; Borges et al., 2001), the entity-relationship model approach (Parent et al., 1999; Parent et al., 2005), and the ontology-based approach. Among these solutions, the ontology-based approach is considered as an approach that can enrich a conceptual schema and thus to improve GDI (Bishr, 1998; Brodeur et al., 2003; Budak et al., 2006; Fonseca et al., 2003; Fonseca et al., 2002; Guan et al., 2003; Hakimpour, 2003; Hakimpour & Geppert, 2002; Hart, et al., 2007; Peachavanish & Karimi, 2007; Souza et al., 2006; Tomai et al., 2007; Uitermark et al., 1999; Visser et al., 2002; Yunpeng et al., 2006) and allow data query and retrieval (Zhang, 2005; Zhao, et al., 2008). In the next section, ontology and how to use ontology to overcome semantic and schema heterogeneity are explored.

2.5 Summary

This chapter is a literature review that examined the definition and attributes of the three facets of geospatial data heterogeneity (i.e., semantic, schematic, and syntactic heterogeneity) and then examined the main solutions to overcome these heterogeneities. Chapter 3 examines the ontology-based geospatial data integration approach introduced in the literature.

CHAPTER 3

An Ontology-based Approach to Geospatial Data Integration (GDI)

The purpose of GDI is to make the data's semantic meaning clear and transparent. The ontology-based geospatial data integration adopts ontology to depict context information, and further to integrate heterogeneous application schemas to accomplish diverse geographic information integration logically.

Ontology originally is a branch of Philosophy which deals with questions about what entities exist or can be said to exist. Ontology in computer science and information science is an explicit formal specification or representation of a set of concepts within a domain that defines the relationships between those concepts. In the GICs, ontology is defined as a collection of shared and formally defined concepts and their relationships (Uitermark et al., 1999; Brodeur et al., 2003; Fonseca et al., 2003).

3.1 Ontology Definition

The OGC defines ontology as a controlled vocabulary that specifies a conceptualisation of a knowledge domain (OGC, 2006). An ontology should meet the following four conditions (Uitermark, 2001):

1) An ontology is a collection of concepts.
2) Concepts should be unambiguously defined.
3) The collection of concepts should be limitative.
4) The collection of concepts should be structured via defined relationships.

Ontology makes concepts clear and defines constraints and the relationships between the concepts. Moreover, ontology provides contextual information for data semantics that aid in characterizing the situation of an entity. Therefore, ontology is a logical foundation for semantic integration because data derive their semantics from ontology and more importantly, ontology provides the computer with an understandable format for the meaning of the data which makes it possible to integrate heterogeneous geospatial data based on their semantic value regardless of their representation.

Fonseca (2002, 2003) and Brodeur et al. (2003) introduced three levels of ontology to geographic information integration. The three levels of ontology are as follows:

1) global ontology (or formal level (Fonseca et al., 2003), upper ontology (Souza, et al., 2006) , and top-level ontology (Fonseca et al., 2002)),
2) domain ontology (or middle ontology (Souza et al., 2006)),
3) and application ontology (or specific ontology (Souza et al., 2006, Guarino's study as cited in Brodeur et al., 2003; Fonseca et al., 2003; Fonseca et al., 2002).

Global ontology is independent of any specific domain, compiling highly abstract concepts where the generic concepts are used to express concept models (Brodeur et al., 2003; Fonseca et al., 2003; Fonseca et al., 2002). *Domain ontology*, as an instance of global ontology, includes an inventory of concepts or conceptual representations which are accepted and shared within an information community (e.g., topographic mapping domain) (Brodeur et al., 2003; Fonseca et al., 2003; Fonseca et al., 2002). *Application ontology* defines and organizes a list of concepts or conceptual representations specific to an application (Brodeur et al., 2003; Fonseca et al., 2003; Fonseca et al., 2002). In the hierarchic structure of ontology, the lower-level ontology is more detailed and must be created based on a higher-level ontology, which means that the lower-level ontology incorporates or inherits the knowledge presented in the higher-level ontology (Fonseca et al., 2002). In the context of geospatial data integration, the application ontology is the ontology for each geospatial dataset. The domain ontology is the vocabulary shared by all the geospatial data that needs to be integrated.

3.2 Ontology-based Geospatial Data Integration

A real-world object is represented in a dataset for a specific purpose and with a defined data model and level of detail which means that objects in reality are abstracted in a specified context. For example, an object in reality is recorded as a Road ("Bank", 3000, 4). How is the meaning of this record deduced? Does the record mean **Road (road name, road length, the number of lanes)** or **Road (road name, the number of buildings on the road, road width)**? As such, one must first to determine the specified context of the information (i.e., the ontology) to arrive at the exact meaning of the records. In this example, one needs to understand the data model, and the meaning of the Road class, the properties of the class and accordingly the property values (i.e., "Bank", 3000, 4). Furthermore, the same object in reality can be represented in another dataset (provided by a different authority) in the form of Street ("Bank", 1, 120). Do these two records refer to the same real-world entity?

The key point of integrating geospatial data is to find relevant objects in different data sources covering different or similar geographic space. Relevant objects that are either disjunctive or coincident in space are objects with similar semantics from different datasets. Finding semantically similar objects is akin to a communication procedure that is based on a common language and shared conceptions that are explicitly represented by ontologies. Based on the above, the idea of ontology-based geospatial data integration is as follows: one first need to explore the ontology (context information) of geospatial datasets that need to be integrated, which are application ontologies; subsequently one needs to match these application ontologies. Matching ontologies is called semantic integration, ontology integration, or ontology mapping. Finally, one can match the heterogeneous classes and properties based on the integrated ontologies and arrive at an (logical) integrated schema, which is named schema matching or schema mapping, or schema integration, respectively. In this example, some relationships need to be confirmed, such as ***Road*** vs. ***Street***. Because of the integration, these two heterogeneous datasets are logically connected and integrated. In an integrated state, users can retrieve the information they want no matter if they search using the keyword "roads" or "streets".

3.2.1 Ontology Exploration

There are numerous ways to obtain ontologies. As for the global ontology, Zhou et al. (2006) adopted the ontology published by Protege of Standford University on GML3.0 (Geography Markup Language) and represented by OWL (Web Ontology Language) as a global ontology. Domain ontologies can be the standards or specifications adopted in a domain or a specific field: Uitermark et al. (1999) used the Geo-Information Terrain Model as the domain ontology for their application. Also, the DHS GDM can be used as a domain ontology for the application of the seven GDC framework fields. The application ontology can be extracted from application schemas which accompany most datasets, and in addition, metadata is also a good reference for creating an application ontology. There are some well-developed software designed for ontology editing, for example the Protege ontology editor (http://protege.stanford.edu/) and the differential ontology editor (http://homepages.cwi.nl/~troncy/DOE/).

3.2.2 Schema Mapping

Schema mapping or matching, the base of schema integration, requires mapping corresponding semantics between the elements of two datasets and includes the classes and properties from the two different schemas (Rahm & Bernstein, 2001, Madhavan et al., 2001). The base of schema matching is semantic integration or matching and is based on what the features actually represent in the real world and not on the way they are represented within the geospatial datasets (Bishr, 1998). Based on ontologies that provide contextual information to understand the semantic meaning of the data in a dataset, semantic relationships between entities stored in different datasets can be mapped (i.e., ontology integration) to facilitate schema integration and finally achieve geospatial data integration.

The ontologies used to explicate the data semantics and achieve semantic integration can be organized in three different ways (Figure 3.1) (Wache, et al., 2001): multiple ontology approaches, single ontology approaches (Comber et al., 2004; Pundt & Bishr, 2002), and hybrid approaches (Hakimpour, 2003; Lutz & Klien, 2006; Stoimenov et al., 2006; Visser, 2004; Zhou et al., 2006). In multiple ontology approaches, each data source has its own application ontology and lacks a common and shared vocabulary and thus makes the comparison of different application ontologies difficult. When applying the multiple ontology approach, inter-ontology mapping is required to compare application ontologies. Alternatively, the single ontology approach adopts a global ontology as a shared vocabulary through which diverse application ontologies are mapped. This approach is suitable for integration of datasets referring to similar product catalogues. While, Wache et al. (2001) argued that it is difficult to build up a global ontology for the datasets referring to different catalogues, and this approach is susceptible to changes in the data sources which can affect the conceptualization of the domain represented in the ontology. The hybrid approach is a mixture of multiple ontology approaches and single ontology approaches. Similar to the multiple ontology approach, each data source has its original application ontology, but a global ontology is employed to make the various application ontologies communicate and map like the single ontology approach. The drawback of this approach is that the existing ontology is hard to reuse since all source ontologies need to refer to the shared vocabulary. Table 1 shows the benefits and drawbacks of the different ontology-based integration approaches.

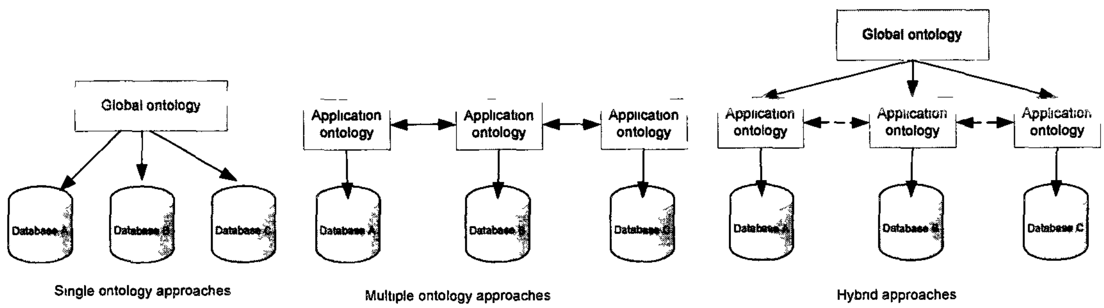

Figure 3.1: Three different organizing ways of using ontology for geospatial data integration (Adapted from Wache, et al., 2001).

Table 3.1: The benefits and drawbacks of the different ontology-based integration approaches

	Single ontology approaches	**Multiple ontology approaches**	**Hybrid ontology approaches**
Implementation effort	Straight-forward	Costly	Reasonable
Semantic heterogeneity	Similar view of a domain	Supports heterogeneous views	Supports heterogeneous views
Adding/removing of sources	Need for some adaptation in the global ontology	Providing a new source ontology; relating to other ontologies	Providing a new source ontology
Comparing of multiple ontologies	--	Difficult because of the lack of a common vocabulary	Simple because ontologies use a common vocabulary

Semantic integration can be achieved through the creation of a global ontology or various matched application ontologies based on, for example, a revised domain ontology. According to the integrated harmonized semantics, schema matching is executed. Zhou et al. (2006) developed an ontology-based GML schema matching methodology (Figure 3.2) based on Rahm and Bernstein's research.

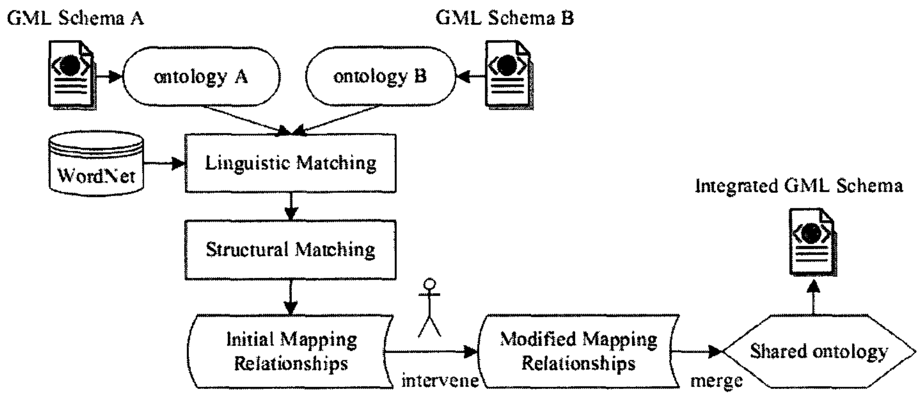

Figure 3.2: The matching is a schema-level matcher approach, performed from linguistic matching of synonyms based on auxiliary information such as WordNet dictionary, then transforming the concepts of the ontology to a concept tree to execute further bottom-up structural matching (Adapted from Zhou et al., 2006)

Based on the kind of matchers (e.g., match algorithms) used in a schema matching procedure, Rahm and Bernstein (2001) classified the major approaches for schema matching as follows:

1) *Instance vs. schema:* matching based on instance data (e.g., data contents) or only on schema-level information;
2) *Element vs. structure matching:* matching is done for individual schema elements, such as attributes (e.g., properties), or for combinations of schema elements;
3) *Language vs. constraint:* matching using a linguistic-based approach, for example based on names or textual descriptions of schema elements, or a constraint-based approach where the constraint may be keys or relationship between the elements;
4) *Matching cardinality:* the matching result may be one or more elements in one schema mapped to one or more elements in another schema, which means the matching cardinality may be 1:1, 1:*n*, *n*:1, or *n*:*m*;

5) *Auxiliary information:* matching not only relies on the input schemas, but also on auxiliary information, such as dictionaries (for example WordNet), global schemas, or previous matching result, and the like.

As described above, one or more elements in one schema are mapped to one or more elements in another schema. Matching relationships at the ontology level or schema level can be further defined (Cruz et al., 2007; Fonseca et al., 2003; Uitermark et al., 2005; Zhou et al., 2006). One example of a mapping relationship (Table 3.2) is matching elements from an application schema to elements from a global schema (Cruz et al., 2007). Based on the matched ontologies and schemas, an integrated schema can be derived. For example, Devogele, et al. (1998) adopted a federated spatial database approach to integrate two road network datasets whose schemas are shown using UML diagrams in Figure 3.3 along with the integrated schema. The same task can be accomplished by using the ontology-based integration approach once the application ontologies can be rectified.

Table 3.2: Possible mapping relationships between an application schema and a global schema

Relationship names	***Description***
Exact	an element in a global schema is equivalent to an element in an application schema
Subset	A relation between a set of elements in a global schema and a single element in an application schema, where each element in the global schema is a subset of the element in the application schema.
Subset complete	A subset relationship, where a set of elements in a global schema is equivalent to the single element in the application schema
Superset	A relationship between a single element in a global schema and a set of concepts in a application schema, where the element in the global schema is a superset of each element in the application schema
Superset complete	A superset relationship, where a element in a global schema is equivalent to a set of elements in the application schema
Comparative	An element or a set of elements in a global schema is mapped to an element or a set of elements in an application schema; While such mapping does not fall in any of the above categories because the elements are more loosely related.

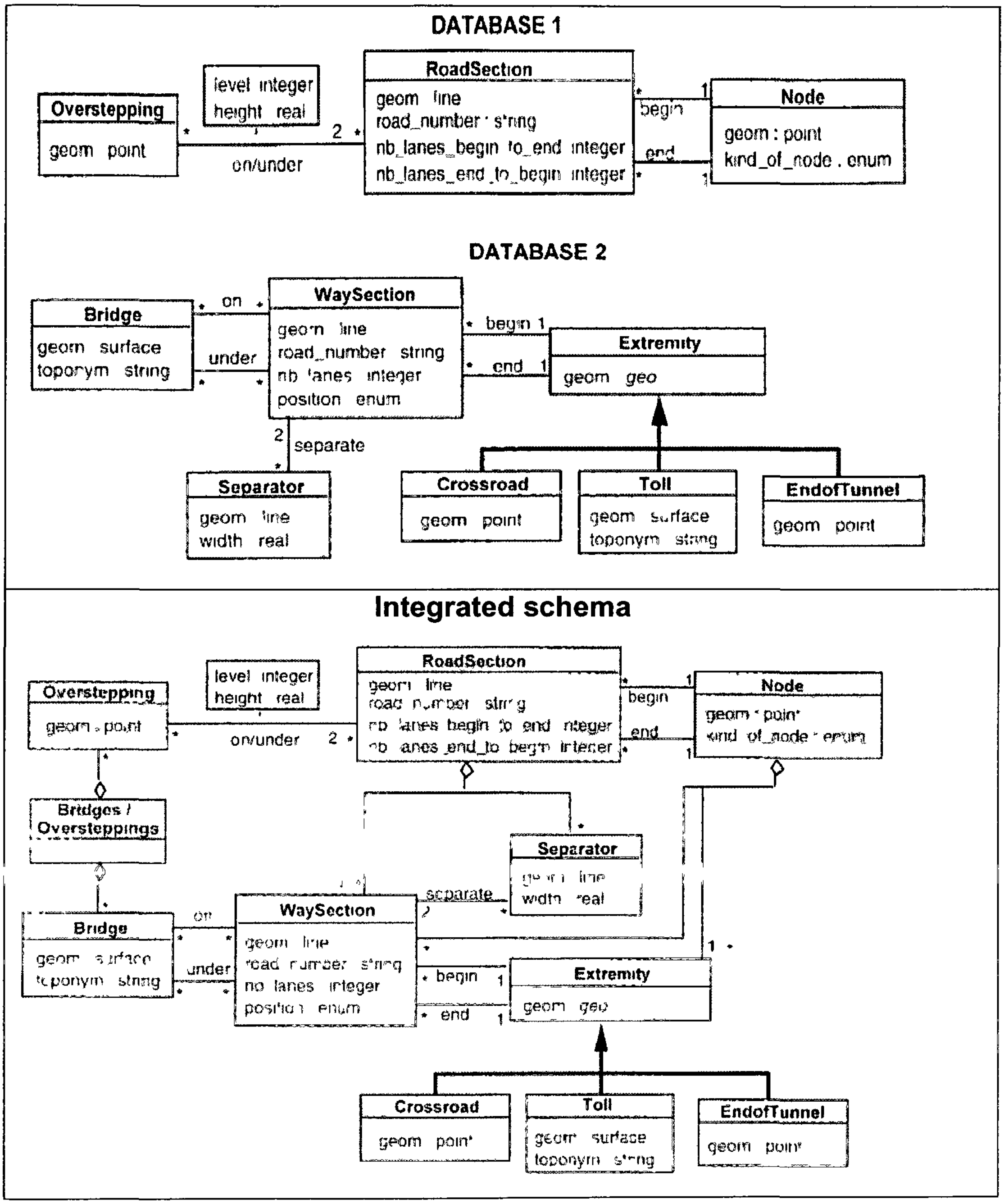

Figure 3.3: The schemas of database 1 and database 2 representing the data sources to be integrated to a schema with a global view. In the UML diagram for the integrated schema, the classes filled with green color are from database 1, while the classes filled with yellow color are from database 2 (from Devogele et al., 1998).

3.3 Summary

This chapter presented a definition of ontology and the three levels of ontology and then outlined the ontology-based geospatial data integration methodology introduced in the literature. The two main steps in the methodology were discussed: how to explore ontologies and how to achieve schema mapping. Three different approaches to organize ontologies and to map schemas were presented: the single ontology approach, the multiple ontology approach, and the hybrid ontology approach. Finally, schema mapping was introduced. The next chapter presents the methodology designed in the thesis research which follows a hybrid ontology approach.

CHAPTER 4

Methodology Development

This research aims to develop an ontology-based methodology for GDI. Illustrated in Figure 4.1, the integration includes application schema development, application ontology exploration, an ontology integration to derive the reference ontology, schema matching, and schema integration. The basis of this methodology is to use ontology to explore data semantics and to find semantically similar objects residing in different data sources. The main steps are explored in this section.

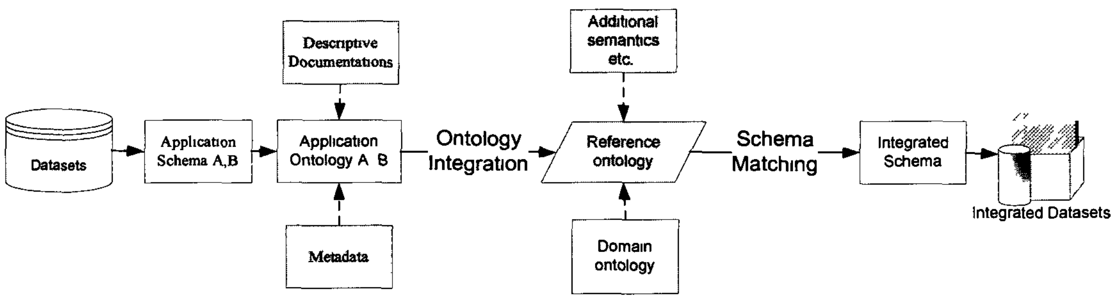

Figure 4.1: An ontology-based geospatial data integration process.

4.1. The Development of Application Schemas and Application Ontology

An application schema shows how the real world is abstracted and represented in a dataset. Some application schemas are provided with data sources. For example, some geospatial datasets are provided as shapefiles and their application schemas can be extracted from the accompanying metadata documents - if they accompany the geospatial data. Some data authorities provide users with access to auxiliary documentation to create application schemas and ontologies. For instance, Geoconnections provides users with such documentation describing the conceptual data model, feature catalogue, and other data description documents for the Canada national road network. In this research, UML (Unified Model Language) is used to represent application schemas because UML is the *de facto* global data modeling standard. Figure 4.2 is an example of road network schema for the Canadian road network data in UML. The road network is composed of road segments, road junctions, toll points, block passage, and ferry connection segments (Figure 4.2).

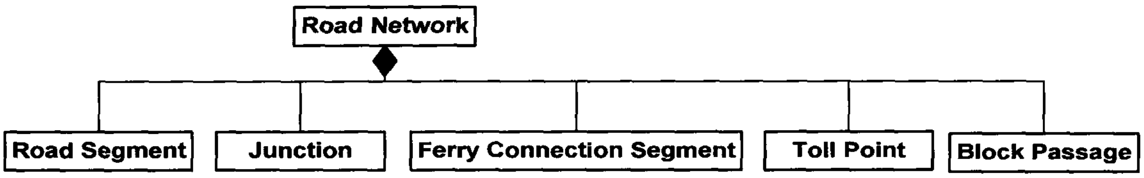

Figure 4.2: A UML road network schema. The black diamond signifies that the second tier of objects are compositionally related, that is to say, the Road Network is composed of the objects in the lower hierarchy.

Like an application schema, each application ontology is specific to one geospatial dataset. The application ontology provides the semantic information for a dataset by defining the explicit meanings of the elements in the application schema and their relationships. The relationships between the elements are mainly inheritance, composition, and association. One needs to examine the accompanying application schema, the metadata document and some other descriptive documents to construct the application ontology. When the existing documentation is incomplete, one also needs to examine the original maps to determine the meaning of the elements and their relationships in the dataset. Figure 4.3 shows the classification of the Canadian road data downloaded from www.Geobase.ca and makes explicit the various values that Road Segment can be assigned.

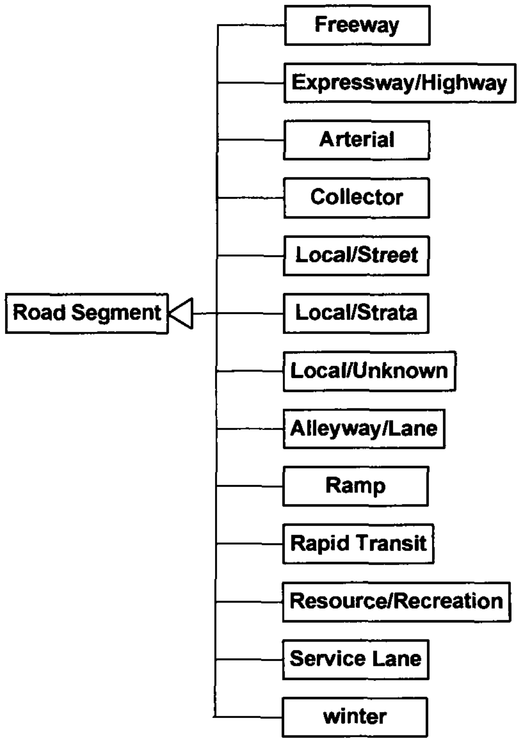

Figure 4.3: The classification of the Canadian Road Segment class.

4.2 The Development of the Domain Ontology for Cross-border Geospatial Data Integration

Geoconnections of Canada and the FGDC of the USA work together to standardize geospatial data and services between these two countries, for example, the FGDC metadata standard is endorsed by Geoconnections and can facilitate geospatial data sharing between these two countries. In this research the Department of Homeland Security Geospatial Data Model (DHS GDM) produced by the FGDC, which is designed to support geospatial interoperability and information sharing for the purpose of homeland security and disaster management (FGDC, 2009), is adopted as the domain ontology to create the reference ontology and integrated schema. The DHS GDM provides specified logical data models for different data authorities to ensure consistency of information. The data model, which is an adaptation of the FGDC model, covers the seven FGDC framework layers: Cadastral, Elevation, GeodeticControl, GovernmentalUnits, Hydrography, Orthoimagery and Transportation, and adds new framework classes for BaseMap, Environmental and Infrastructure, Pipelines and Multi-modal Routes (FGDC, 2008). Unfortunately, the definition or description (i.e., ontology) of the feature subtypes is not included in the model. In some cases, like this research, in order to get the semantic meaning of a road classification, users may need to redirect to the documents provided by the official partners of the Department of Homeland Security, for example, the U.S. Department of Transportation is responsible for the road datasets in the DHS GDM.

WordNet (http://wordnet.princeton.edu/) is a useful tool to aid the development of a reference ontology. WordNet is a lexical database that provides synonyms for any english word. Moreover, WordNet specifies the semantic relations between these synonym sets. For example, if you search for the synonym words for "Road", it provides hypernyms (road is a kind of ...), hyponyms (...is a kind of road), holonyms (road is a part of ...), and meronyms (...is a part of road) nouns related to "Road" as well as the short and general definitions for each related noun. Therefore, WordNet can be interpreted as a lexical ontology in this research. This can be particularly useful in developing a non-English reference ontology when incorporating English datasets.

In this research, the DHS GDM, together with WordNet, is used as a starting point for the construction of a reference ontology.

4.3 The Construction of the Reference Ontology

As described above, the semantic heterogeneity presented in various datasets is the main obstacle for data integration. Therefore, semantic reconciliation for production of application ontologies is a prerequisite for geospatial data integration. The approach for ontological integration in this research will be a hybrid approach where application ontologies and a reference ontology (which acts as a domain ontology) will participate in the integration procedure.

The reference ontology acts as a bridge to match the concepts from different application ontologies. Directly capturing the semantic relationships between different application ontologies is difficult, especially when integrating several application ontologies whose datasets do not describe the same geographic space and are from different data authorities. The application ontologies from these different datasets are all mapped to the reference ontology. Therefore, concept matching between these application ontologies is thereby achieved. The reference ontology is based on a domain ontology in a discipline, and then refines the concepts in the domain ontology. A reference ontology is built in the following way: if object classes in a reference ontology are as a subset of object classes from a domain ontology, object classes from this subset are refined into subclasses in a taxonomic structure; if object classes from different application ontologies are composed of each other, the added object classes will be structured in a partonomy classification (Uitermark, 2001).

4.3.1 Object Classes in the Example Application Ontologies

A good example illustrating how to build a reference ontology can be based on the methodology developed by Uitermark et al. (2005) and Stoimenov and Djordjevic-Kajan (2005). The Canada road segment data (dataset A) from www.geobase.ca and the USA road segment data (dataset B) from 2008 TIGER/Line data are used in this example. The local street and lower level street classes for road segments in dataset A and B are shown in Table 4.1 and 4.2, respectively.

Table 4.1: The local street and lower level street classes for road segments in dataset A, the Canadian Road Segment dataset.

Object Class Name	***Application ontology concept definition***
Local / Street	A low-speed thoroughfare dedicated to provide full access to the front of properties.
Local / Strata	A low-speed thoroughfare dedicated to provide access to properties with potential public restriction such as: trailer parks, First Nations, strata, private estates, seasonal residences.
Local/ Unknown	A low-speed thoroughfare dedicated to provide access to the front of properties but for which the access regulations are unknown.
Alleyway/ Lane	A low-speed thoroughfare dedicated to provide access to the rear of properties.
Ramp	A system of interconnecting roadways providing for the controlled movement between two or more roadways.
Resource/ Recreation	A narrow passage whose primary function is to provide access for resource extraction and may also have served in providing public access to the backcountry.
Rapid Transit	A thoroughfare restricted to public transit buses.
Service Lane	A stretch of road permitting vehicles to come to a stop along a freeway or highway. Such as Scale, service lane, emergency lane, lookout, and rest area.
Winter	A road that is only useable during the winter when conditions allow for passage over lakes, rivers, and wetlands.

Table 4.2: The local street and lower level street classes for road segments in dataset B, the USA Tiger dataset.

Object Class Name	***Feature class***	***Application ontology concept definition***
S1400	***Local neighbourhood road, rural road, city road***	Generally a paved non-arterial street, road, or byway that usually has a single lane of traffic in each direction. Roads in this feature class may be privately or publicly maintained. Scenic park roads would be included in this feature class, as would (depending on the region of the country) some unpaved roads.
S1500	***Vehicular Trail (4WD)***	An unpaved dirt trail where a four-wheel drive vehicle is required. These vehicular trails are found almost exclusively in very rural areas. Minor, unpaved roads usable by ordinary cars and trucks belong in the S1400 category.
S1630	***Ramp***	A road that allows controlled access from adjacent roads onto a limited access highway, often in the form of a cloverleaf interchange. These roads are unaddressable.
S1640	***Service drive usually along a limited access highway***	A road, usually paralleling a limited access highway, that provides access to structures along the highway. These roads can be named and may intersect with other roads.
S1710	***Walkway/Pedestrian Trail***	A path that is used for walking, being either too narrow for or legally restricted from vehicular traffic.
S1720	***Stairway***	A pedestrian passageway from one level to another by a

		series of steps.
S1730	***Alley***	A service road that does not generally have associated addressed structures and is usually unnamed. It is located at the rear of buildings and properties and is used for deliveries.
S1740	***Private Road for service vehicles (logging, oil fields, ranches, etc.)***	A road within private property that is privately maintained for service, extractive, or other purposes. These roads are often unnamed.
S1780	***Parking lot road***	The main travel route for vehicles through a paved parking area.
S1820	***Bike path or trail***	A path that is used for manual or small, motorized bicycles, being either too narrow for or legally restricted from vehicular traffic.
S1830	***Bridle path***	A path that is used for horses, being either too narrow for or legally restricted from vehicular traffic.

4.3.2 Object Classes in the Domain Ontology and the Reference Ontology

In order to match the concepts residing in different application ontologies, the concepts in a reference ontology need to be detailed enough to cover all the information from the datasets to be integrated. Therefore, the concepts coming from the domain ontology need to be refined and redefined. In the DHS GDM, roads are classified into ***Highway***, ***Arterial road***, ***Collector road***, and ***Local road***. The lowest level of road classes is ***Local road*** which is too generalized for matching various concepts in existing application ontologies in dataset A and B (shown in Table 4.3).

Table 4.3: Concepts in the domain ontology.

Object Class Name	***domain ontology concept definition***
LocalRoad	Consists of all roads not defined as arterials or collectors; provide limited mobility and is the primary access to residential areas, businesses, farms, and other local areas.

For Dataset A, the local road concept is refined into three subclasses: ***Local/Street***, ***Local/Strata***, and ***Local/Unknown***. However, these subclasses are not the lowest classes in the road hierarchy in the Dataset A ontology. There are other classes defined for road segments, such as ***Alleyway/Land***, ***Ramp***, ***Resource/Recreation***, ***Rapid Transit***, ***Service Lane***, and ***Winter Road***. Meanwhile, for dataset B, the local road concept is refined into class ***S1400*** which means Local Neighbourhood Road, Rural Road, and City Road. Except for ***S1400***, dataset B also defines other road classes such as ***Ramp***, ***Alley***, etc. which are some kinds of the ***LocalRoad*** class based on the definition provided by DHS GDM. Therefore, the class ***LocalRoad*** in the domain ontology has to be refined in order to provide a basis for application ontology integration. In this research, ***LocalRoad*** class can be sub-classified into ***local road***, ***service road***, and ***other passage*** classes. Table 4.4 lists the concepts in the reference ontology for these example data.

Table 4.4: Concepts in the domain ontology and the reference ontology that can embody the concepts in Dataset A and B.

Domain ontology concept	***Refined subclass in the reference ontology***	***Definition of refined subclass in the reference ontology***
LocalRoad	***Local road***	Generally a non-arterial street, road, or byway that usually is a low-speed thoroughfare dedicated to provide full access to the front of properties (or addressed structures).

	Service road	A service road that does not generally have associated addressed structures and is usually unnamed.
	Other road	A narrow path or passage which usually is not accessed by vehicle.

4.3.3 The Reference Ontology

The definition and classification of object classes in the reference ontology are based on the application datasets to be integrated. The fundamental relationships between the concepts in a reference ontology are generalization/specialization (Is-a, sub/super, or Inheritance) and aggregation (component/composite, whole-part). Concepts are grouped into a taxonomy with superclasses and subclasses via generazation/specialization relationships or into a partonomy with composite and component classes via aggregation relationship. Therefore, concepts in the reference ontology are organized into a hierarchical tree-like structure (Figure 4.4), that will act as a bridge to communicate the concepts from diverse application ontologies.

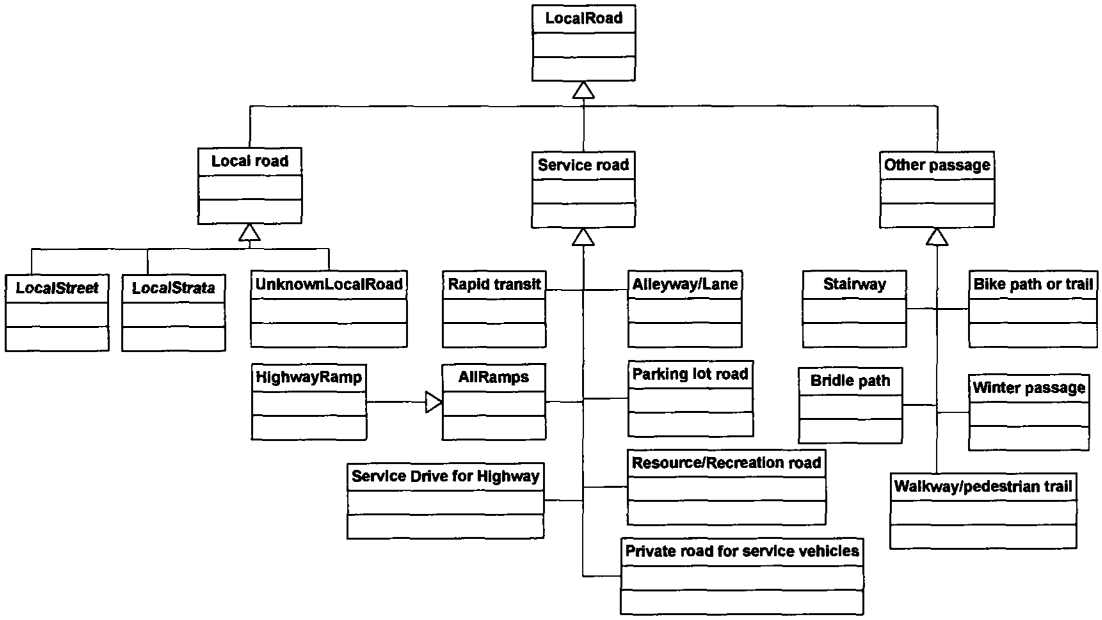

Figure 4.4: The reference ontology for Local Roads shown in UML. It also can be expressed in XML(Extensible Markup Language), RDF(Resource Description Framework), and UML format.

4.4 Semantic Integration

In order to facilitate the integration of ontology semantics, four semantic relationships, which are refer-to, semantic equivalence, semantic generalization, and semantic aggregation relationships, are defined in this thesis research.

4.4.1 Refer-to Relationship

A semantic relationship, named refer-to, is defined to communicate the concepts from the constructed reference ontology and diverse application ontologies (Table 4.5). A constraint for this relationship is: for a given reference ontology concept ***a***, there is at most one application concept that satisfies the relationship refer-to, but for a given application concept ***b***, there may be more reference ontology concepts satisfying this kind of relationship (Uitermark, 2001). This constraint is used to make sure that the reference ontology is refined enough to express every semantic similarity between concepts from the reference ontology and application ontologies.

Table 4.5: The relationship between the concepts from a reference ontology and an application ontology.

Name	***Definition***	***Examples***
Refer-to (***a***, ***b***)	Concept ***b*** in an application ontology is refer-to concept ***a*** in a reference ontology if concept ***b*** is semantically similar to concept ***a***	refer-to (local road, *S1400*) refer-to (AllRamps, Ramp in Canada database)

4.4.2 Semantic Relationships Among Application Ontologies

In order to achieve ontology integration, the semantic relationships between concepts from different application ontologies can be defined as semantic equivalence, semantic generalization, and semantic aggregation. The relationships are defined as follows(Uitermark, 2001).

1) Semantic equivalence (SREquivalent): concept ***b*** and ***c*** are semantically equivalent if there exists a concept ***a***, and both concepts ***b*** and ***c*** are all refer-to ***a***, recorded as SREquivalent (***b***, ***c***). Concept ***b*** and ***c*** actually represent the same phenomena in reality. This relationship maps as a 1:1 relation among concepts.
2) Semantic generalization (SRGeneration*)*: concepts ***b*** and ***c*** are semantically inherited if there exists concept ***a1*** and ***a2*** in a reference ontology, concept ***b*** and ***c*** are refer-to ***a2*** and ***a1*****,** respectively, and ***a1*** is a subclass of ***a2***. This relationship is recorded as SRGeneration (***c***, ***b***). Being in this kind of relationship, concept ***c*** is a subset of concept ***b***, which means concept ***c*** not only inherits the attributes and methods of concept ***b*****,** but also adds its own special attributes and methods. Here, the inheritance relationship may be more than two levels.
3) Semantic aggregation (SRAggregation): concept ***b*** and ***c*** are have semantic composition if there exist concept ***a1*** and ***a2***, concept ***b*** and ***c*** are refer-to ***a1*** and ***a2***, respectively, and ***a1*** is a component of ***a2***. This relationship is recorded as SRAggregation (***b***, ***c***). Being in this relationship, concept ***b*** is a part of concept ***c***.

Based on the defined relationships above, an integrated ontology can be derived (Figure 4.5).

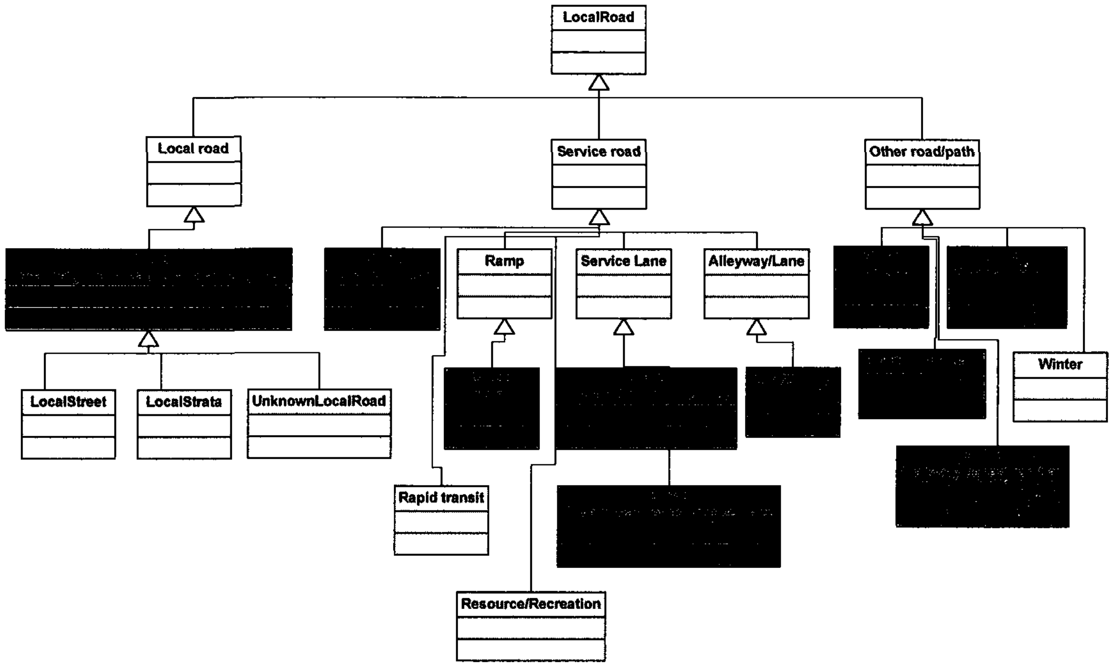

Figure 4.5: The semantic relationships between Dataset A and B. The concepts with green backgrounds are from Dataset A; the concepts with blue backgrounds are from Dataset B.

4.5 Schema Matching and Integration

In this research, schema matching will be schema-level matching, which means that detailed information at the instance level will not be considered. Because the elements (such as names, description, data type, relationships, constraints, and schema structure) in an application schema are organized in the accompanying application ontology, the integrated schema will be achieved based on the integrated ontologies. In this example, since the concepts in the application ontologies are the same as that in application schemas, the integrated schema is same as the integrated ontology shown in Figure 4.5.

4.6 The Use of Semantic Identifiers

The conventional way to represent objects in GICs is a 3-tuple {Object identifier, Geomteric attributes, Thematic attributes}, which can be simply recorded as {ID, G, T}. Inspired by the study of Bishr (1998), in this research, geographic objects are represented with a 4-tuple which is {ID, Geometric attributes, Thematic attributes, Object semantics}, shown in Figure 4.6. Each object class will have a unique semantic descriptor acting in a role to distinguish homonymous concepts or to identify synonymous concepts from various application ontolgoies. Accordingly, all the instances belonging to the same object class will have the same semantic descriptor. To avoid data redundancy, all the semantic descriptors will be recorded in tables which are designed to trace the semantic relationships among the concepts.

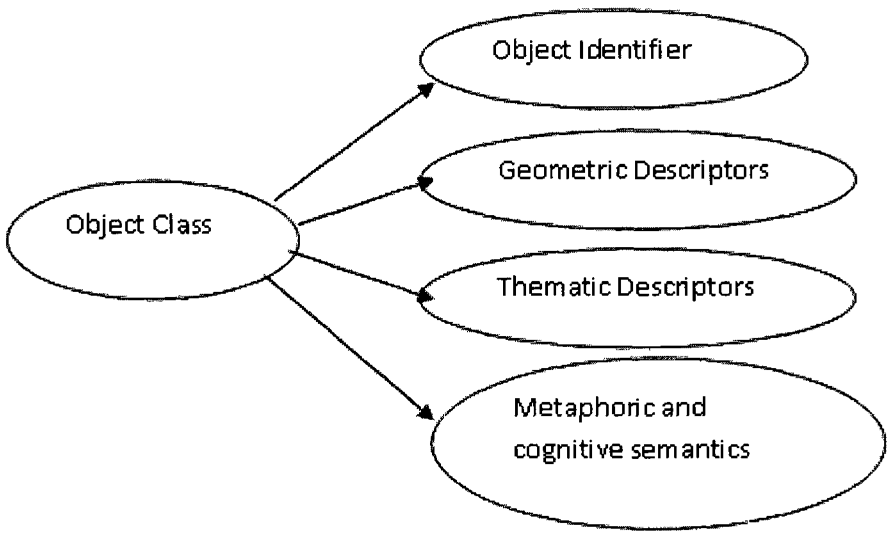

Figure 4.6: Adding semantics to the conventional representation of geospatial objects. (Bishr, 1998).

Within the ontology-based integration approach, the classes from the geospatial datasets to be integrated are linked via the relationships of SREquivalence, SRGeneration, or

SRAggregation. Therefore, an integrated schema for the datasets to be integrated will be derived and will form the basis of data integration and data query.

The ontology-based integration method can overcome the problems resulting from multi-scale and multi-representation issues. A larger-scale dataset is prone to have more fragments to represent than in a smaller-scale dataset due to the different scales that intrinsically imply different levels of aggregation. The fragments in the larger-scale dataset will be a component of the objects in the small-scale dataset, which means a semantic aggregation relationship exists between these classes. Therefore, the ontology-based integration approach can be applied across scale-varying datasets.

4.7 The Conceptual Framework of the Ontology-Based Geographic Information Integration

The ontology-based geospatial data integration approach can be depicted in Figure 4.7. Datasets within the GIS domain are abstracted from the real world. The real world phenomena are categorized by abstraction rules into classes with a taxonomy and partonomy structure. These classes are defined as object classes in a conceptual world and documented in a domain ontology (the abstract procedure presented in Figure 2.1). Dataset A and B are the heterogeneous geospatial datasets to be integrated. Both of them have their own concepts which are defined and documented in the application ontology A and B, respectively. The reference ontology is constructed based on the domain ontology, application ontologies, and other knowledge such as WordNet, and maybe surveying rules. The corresponding object classes in different datasets are defined via the use of a refer-to semantic similarity relationship. Therefore, the heterogeneous geospatial datasets are logically integrated with the aid of the construction of different levels of ontologies.

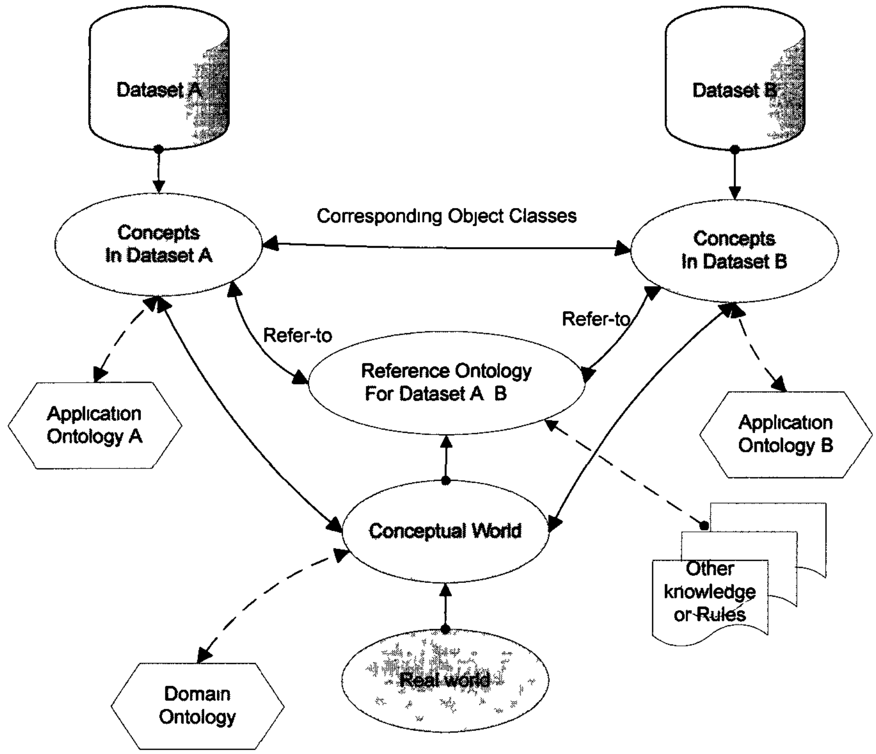

Figure 4.7: An ontology-based framework for geospatial data integration (adapted from Uitermark, 2001).

4.8 Summary

This chapter introduced the methodology designed for horizontal geospatial data integration by using the example of the integration of road segment data from Canada and the USA. This chapter began with the introduction of the integration process, and then explained the detail of each step in the integration process which includes the development of application schemas and application ontologies, the construction of the domain ontologies, the reference model and schema integration. In addition, four kinds of semantic relationships are defined (i.e., refer-to; semantic equivalence, semantic generalization, and semantic aggregation) and are used to aid the integration of ontologies. The last part of this chapter presented the conceptual framework of the ontology-based geographic information integration. The next chapter introduces how to implement geospatial data integration in the context of Canada-US bordering areas based on the designed methodology introduced in this chapter.

CHAPTER 5

The Implementation of Cross-border Geospatial Data Integration

Geospatial data created by Canada and the USA follow the universally adopted standards and specifications of the ISO and the OGC. However, such standards do not specify semantic and schema homogeneity. As such, individual datasets in Canada and the USA have semantic and schema differences because of different political, cultural, and technological approaches as well as differences in the intended use of their respective data. As such these data are difficult to use in seamless cross-boundary research. Reiterating a previous example makes this clear: the example shown in Figure 1.3 shows the boundaries of the US states (named dataset1) and Canadian provinces (named dataset2) and *prima-facia* they look well matched; however, if users want to retrieve Michigan state and the province of Ontario, they need to know by examining the metadata that the fields of ***Name*** in dataset2 and ***State*** in dataset1 refer to the name of a state/province. The user would then need to execute two separate SQL (Structured Query Language) statements: "select * from dataset1 where state='Michigan'" and "select * from dataset2 where name='Ontario'". The results are two separate datasets. However, if users prefer one dataset for further application they can **Union** these two retrieved datasets into one dataset which includes all the attribute fields (e.g., ***Name*** and ***State***) from dataset1 and dataset2; this however, makes the result more complicated.

Based on the cross-border geospatial data integration methodology in Chapter 4, semantically and schematically heterogeneous geographic datasets can be logically integrated using the concepts, attributes, and relationships residing in various geospatial dataset schemas. Integration means that users can freely access and retrieve data from these logically integrated datasets without prior knowledge of various schemas and concepts from the different data sources. Fundamentally then, a layer retrieved from an integrated database (e.g., the first-level of country administration divisions which is province in Canada and state in the US) presents a unified schema that reconciles the heterogeneous semantics and schemas and also facilitates the application of the results. By reference to Figure 1.3 again, once dataset1 and dataset2 are logically integrated, only the field called '***Name***' (for US States or Canadian Provinces) is seen by the user and the fields '***Name***' or '***State***' from dataset1 and dataset2 are not revealed or required.

In this research, framework datasets along the USA and Canada border are integrated based on the integrated methodology. The definition of framework datasets, the schema integration procedure, the interface design, and the development of a MS ACCESS Database for recording ontology relationships are presented in this section.

5.1 Framework Datasets

Geospatial framework data are core data themes that are of critical importance to the spatial data infrastructure (Africa SDI, 2009; FGDC, 2008a; GeoConnections, 2008a). The goal of building national framework datasets is to share data from different sources, to reduce costs, and to improve decision making (FGDC, 2008a; GeoConnections, 2008a). Worldwide there are many initiatives aiming at the development and improvement of National Spatial Information Frameworks (NSIF). Examples of the programs are found on every continent, for example, China (Chen, 1999), South Africa (www.nsif.dla.gov.za, 2009), Canada (GeoConnections, 2008a), USA (FGDC, 2008a), Austria (ANZLIC, 2009), and Europe (Masser, 2000) among others. These geospatial framework data were produced based on universally adopted specifications and standards like ISO 19100 and OGC specifications and standards (FGDC, 2008d; GeoConnections, 2008a). In addition, the development of the Global Spatial Data Infrastructure (GSDI) is ongoing which is aiming to achieve interoperability of digital geographic data and technologies to support decision making at all scales for multiple purposes (Masser, 2003).

Geospatial framework datasets provide contextual and reference information for the area of interest. The FGDC defines seven critical framework themes-Cadastral, Digital orthoimagery, Elevation, Geodetic control points, Governmental unit boundaries, Hydrography, and Transportation (FGDC, 2008d). Based on the FGDC standards, the department of Homeland Security Geospatial Management Office, FGDC, developed the DHS GDM to support geospatial information interoperability and sharing. The DHS GDM added Basemap, Environment, and Infrastructure data themes, and included pipelines into the Transportation theme (FGDC, 2008c). The CGDI also defines the content of Canadian geospatial framework data which includes three principal forms: alignment layers, land feature/form layers, and conceptual layers (Geoconnections, 2008a).

In the context of Canada-US cross-border regions, this research provides the methodology and example of integration of the Canadian geospatial framework data (Geoconnections, 2008a) and the cross-border framework datasets for multi-purpose applications (such as critical infrastructure protection, public health, public safety and security) (shown in Table 5.1):

1) Alignment data: Alignment data include geodetic data which have higher accuracy since they are used to position other geospatial information. Therefore, the alignment data are critical for the reliability and use of other geospatial datasets, which makes them the base of the other geospatial framework data.
2) Land feature data: Land feature data depict reality features which include national road network, electrical transmission lines and pipelines, national hydro network such as rivers and lakes, and manufactured structures such as bridges, airport terminals, ports and dams. These data can be directly used for an application or as a reference for other data.
3) Conceptual data: conceptual data are mainly various jurisdictional boundaries that include geopolitical boundaries of countries, provinces, and municipalities. Geographical names and postal code information were also collected in this research.

The geographic coordinate system of the above datasets is defined as the North American Datum of 1983 (NAD 83).

Table 5.1: The collected framework datasets along the Canada-US border area.

Data Topic	***Dataset Name***	***Source***
Alignment data	Canada geodetic control points	Geoconnections, www.geobase.ca
	The US geodetic control points	USA National Geodetic Survey (NGS) www.ngs.noaa.gov/cgi-bin/datasheet.prl
Land feature data	Canada Road data	Geoconnections, www.geobase.ca
	The USA road dataset	2008 TIGER/Line data from U.S. Census Bureau http://www.census.gov/geo/www/tiger/tgrshp2008/tgrshp2008.html
	Ontario electrical pipelines	Canada DMTI Spatial Inc.
	The USA electrical pipelines	2008 TIGER/Line data from U.S. Census Bureau
	Canada airport data	http://www.plews.ca/AirportsinCanada.htm
	The USA airport data	2008 TIGER/Line data from U.S. Census Bureau
	Canada hydro network datasets	Geoconnections, www.geobase.ca
	The USA hydro data	2008 TIGER/Line data from U.S. Census Bureau
Conceptual data	Canada province boundaries	Canada DMTI Spatial Inc. Geoconnections, www.geobase.ca
	Canada municipality boundaries	Canada DMTI Spatial Inc.
	The USA state boundary dataset	2008 TIGER/Line data from U.S. Census Bureau
	The USA county boundary dataset	2008 TIGER/Line data from U.S. Census Bureau
	Canada postcode dataset	Canada DMTI Spatial Inc.
	The USA zip code dataset	http://www.zip-codes.com/

5.2 Schema Integration

The cross-border framework datasets in this research cover different domains. These domains include geodetic points, political units (province and municipality) boundaries, roads, power pipelines, hydroelectricity and the like. The framework data produced by different countries and organizations are heterogeneous in semantic and schematic aspects, which means the ontology of concepts and schemas residing in the collected datasets vary from Canada to the USA. Logically integrating these datasets via an ontology-based schema integration approach is an efficient way to reuse and share these datasets. The schema integration requires the following steps: first, the construction of a reference model based on the retrieved domain ontology and application ontologies, and second, matching application schemas based on the developed reference model. The domain ontology for each data theme is developed based on the DHS GDM. The application ontologies are retrieved by examining data metadata, data documents, and even maps. The detailed schema integration procedure for geodetic control points, political unit boundaries, and road datasets covering the Canada-US border area are detailed in the following sections.

5.2.1 Geodetic Control Points

The geodetic control point network provides a common reference system for establishing the coordinates of all geographic data. Therefore, geodetic control points play a crucial role in developing all framework data and users' application data (FGDC, 2008d). The main on-ground features of geodetic control points are geodetic control stations which are represented by points in GICs. These often monumented points have precisely measured horizontal and/or vertical coordinates. The horizontal and/or vertical coordinates of the control points and related information which includes the name of the station, unique identification, orthometric height, and ellipsoid height for example are the main components of the ontology of geodetic control point information (FGDC, 2008d).

Based on the Geographic Information Framework Data Content Standard for Geodetic Control Points (FGDC, 2008d), the standard information for each geodetic control point requires four basic elements: designations, coordinates, accuracy, and geodetic datum. The designations element refers to three types of identifiers for each point which include one mandatory element: a unique identifier, and two optional elements: a descriptive identifier and a uniform resource identifier. The element of coordinates includes horizontal coordinates in latitude and longitude format and vertical measure in either orthometric height or ellipsoid height. The accuracy information includes local and network accuracy provided in meters, expressed at the 95% confidence level. The geodetic datum provides reference for horizontal and vertical coordinates. These four basic elements are adopted into the DHS GDM, and also are considered as the domain ontology of geodetic control points in this integration research example.

In this research, one USA control point dataset (scale 1: 1,500,000) was collected and contains a set of geodetic control stations maintained by the US national Geodetic Survey. Each US geodetic control station in this dataset has precisely measured Latitude/Longitude coordinates and Orthometric Heights and provides the reference for horizontal and vertical coordinate control throughout the US. Three Canadian control point datasets: the Canadian Base Network (CBN1, scale 1:250,000), Primary Vertical Network (PVBMC1, scale 1:50,000), and Federal 2D Densification Network (FED3DC1, scale 1: 250,000) were likewise collected. The three Canadian datasets were produced for different purposes. The concepts from the control point domain ontology and from the ontologies of these datasets are listed in Table 5.2.

Table 5.2: Concepts in domain ontology and application ontologies of geodetic points.

concepts in control point domain ontology	***related concepts in Canada control point dataset***	***Application ontology descriptor***	***related concepts in the USA control point dataset***	***Application ontology descriptor***
A unique identifier	***uniqueNo***	unique number	***PID***	Permanent identifier assigned by NGS
a descriptive identifier	***name***	station name	***name***	station name
horizontal coordinates-1	***latitude***	latitude in deg-min-sec format(e.g. N46°22′37.0271″)	***latitude***	latitude in decimal degree format(e.g. 47.70917).
horizontal coordinates-2	***longitude***	longitude in deg-min-sec format (e.g. W62°08′08.7689″))	***longitude***	longitude in decimal degree format (e.g. -116.89389)
vertical coordinate-orthometirc elevation	***elevation***	orthometirc height in meters indicating the height above the geoid.	***elevation***	the orthometric height in meters indicating the height above the geoid.
accuracy of horizontal coordinates	***latstddev longstddev***	latitude/longitude standard deviation in meter in the NAD83CSRS system	***pos_order***	order of accuracy for the latitude, longitude, shoul d be one of the following a,b,1,2,3
accuray of vertical coordinate	***ellstddev***	ellipsoidal height standard deviation	***elev_order***	order of accuracy for the elevation, should be one of the following 0,1,2,3
reference for horizontal coordinates	***horizDatum***	horizontal datum	***pos_datum***	datum of the latitude,longitude.
reference for vertical coordinate	***vertDatum***	elevation datum	***elev_datum***	datum of elevation

Special attention should be paid to the differences that are evident in the concepts of the application ontologies. For example, the latitude and longitude values for the geodetic points from US dataset and Canada datasets are in different formats (see the example listed in Table 5.2); The horizontal and vertical datum of US points are NAD83 and NAVD88, respectively; while the datum of Canadian points are NAD83CSRS and CGVD28, respectively. The accuracy systems for US and Canada datasets are different: one is represented by the standard deviation, another is by the pre-defined accuracy order system; moreover, all of these accuracy indicators are different from the domain ontology.

The reference model of the geodetic control point domain includes the following concepts: ***uniqueID***, ***name***, ***latitude***, ***longitude***, ***horizontalReferenceSystem***, ***orthometricHeight***, ***orthometriHeightDatum***, ***horizontalAccuracy***, and ***verticalAccuracy***. The ontology description for each concept in the reference model or application schemas named ontology descriptor is recorded in a database, which will provide context information of the concept when users employ these data. Therefore, the aforementioned differences are recorded as application ontology descriptors. The integrated schema for the geodetic control point datasets using the explicated ontology is shown in Figure 5.1. Users can access and retrieve the concepts presented in the reference model like ***uniqueID***; while the heterogeneous concepts in the different datasets like ***PID*** and ***uniqueNo*** are not public to users.

GeodeticControlPoint
-uniqueID (US.PID;Canada.uniqueNo) : String -name(US.name;Canada.name) : String -latitude(US.latitude;Canada.latitude) : String -longitude(US.longitude;Canada.longitude) : String -horizontalReferenceSystem (US.pos_datum,Canada.horizDatum) : String -orthometricHeight ((US.elevation;Canada.elevation)) : Double -orthometricHeightDatum (US.elev_datum; Canada.vertDatum) : String -horizontalAccuracy(US.pos_order,Canada.latstddev& longstddev) : Double -verticalAccuracy(US.ele_order,Canada.ellstddev) : Double

Figure 5.1: The integrated schema for the geodetic control point datasets. "–uniqueID (US.PID; Canada.uniqueNo): String" means concept ***PID*** from US dataset is Refer-to the standard concept ***uniqueID***, and so is concept ***uniqueNo*** from Canada dataset. Thus concept ***PID*** is SREquivalent to concept ***uniqueNo***. String is the data type of these concepts in the datasets.

5.2.2 Political Unit Boundaries

In this research, under cross-border circumstances, the boundaries of the political entities or units were chosen as the geographic area boundaries. Two levels of political units were encompassed which are the first (province/state) level and the second (municipality/county) level. In the DHS GDM, the required concepts for the first level political unit boundaries are ***stateFIPSCode*** which is a unique identifier for a province/state, and ***population***; the required concept for the second level political unit boundaries is ***detailCity*** which is the name of a municipality/city. Two datasets for Canadian provincial boundaries (scale: 1: 20, 000) and two datasets for US state boundaries (scale: 1:100,000) were collected. The ontology of the concepts retrieved from the US and Canadian province/state datasets and the relationships among these concepts and attributes are listed in Table 5.3. One dataset for the Canadian municipality boundaries (scale: 1:50,000) and one dataset for US counties (scale: 1:100,000) were collected and the ontology of the concepts extracted from these two datasets and the relationships among the concepts and attributes are shown in Table 5.4.

Table 5.3: The concepts in the province boundary domain ontology and province/state datasets. The concept of ***Descriptive Identifier*** is not required in the domain ontology, but for convenience, it is added into the reference model.

Concepts in the domain ontology	***StateFIPSCode***	***Population***	***Descriptive Identifier***
related concepts in Canada province dataset1	***UUID***		***Name***
application ontology descriptor	unique identity number		the name of a province or territory
related concepts in Canada province dataset2		***Pop2001***	***Name***
application ontology descriptor		population in 2001	the name of a province or territory
related concepts in USA state dataset1	***State_FIPS***		***State***
application ontology descriptor	2-digit FIPS code of the state or state equivalent		the name of the state or state equivalent
related concepts in USA state dataset2	***Statefp***		***Name***
application ontology descriptor	2-digit FIPS code of the state or state equivalent		the name of the state or state equivalent

Table 5.4: The concepts in the municipality boundary domain ontology and municipality/county datasets.

concepts in the domain ontology	***related concepts in Canada dataset***	***application ontology descriptor***	***related concepts in US dataset***	***application ontology descriptor***
detailCity	***Name***	the name of a municipality	***County***	the name of a county

Some ontology concepts from the political unit boundary datasets are not presented in the domain ontology, but they are shown in both the Canada and the USA datasets and have the same contextual information. Thus, these concepts (e.g., ***Name***) are adopted into the reference model. The concepts in the reference model of province/state boundaries are ***ID***, ***Name***, ***ProvinceCode***, ***Name_abbreviate***, ***Population***. The concepts in the reference model of municipality/county boundaries are ***Name***, ***ProvinceName***, and ***Area***. The integrated schemas for the first and second level political unit boundaries and their relationships to the application schemas are shown in Figure 5.2.

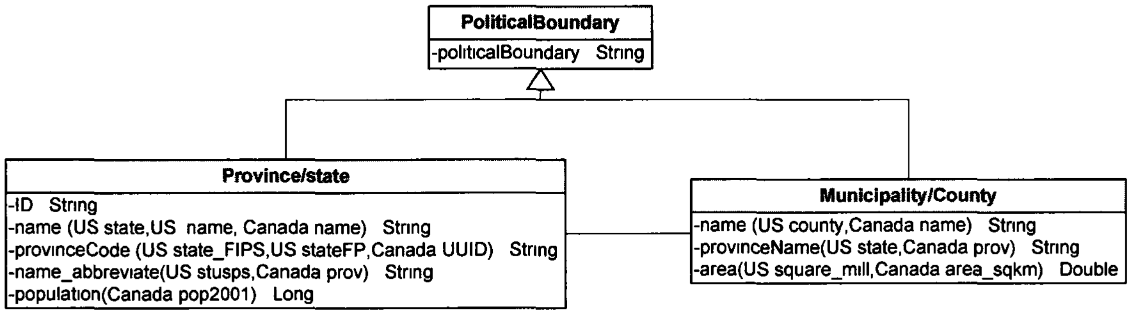

Figure 5.2: The integrated political unit boundary dataset schema. The hierarchy shows that the province/state class and municipality/county class are a kind of political boundary. The province/state class is associated with the municipality/county class via the attribute ***name*** in the province/state class and attribute ***ProvinceName*** in the municipality/county class. Even though some concepts like ***population*** are not present in all the datasets, they are retained in order to add new data sources in the future.

5.2.3 Road

In the Geographic Information Framework Data Content Standard for Transportation Base, the road network is represented by road segments, road points, and road paths (FGDC, 2008d). A road segment represents a linear section of the physical road network designed for human or vehicular movement (FGDC, 2008d). In practice, a road segment can be defined based on the change in the number of lanes, the width of lanes, having traffic volume, etc. A road segment has a start point and an end point that ensure a complete road network system without gaps or overlaps (FGDC, 2008a). Road points provide topological connections between road segments, which could be toll facilities, road intersections, road direction changing points, or traffic light positions. Road paths represent how the road segments are organized and used as administrative routes (FGDC, 2008d) like Highway 415 and Carling Avenue in Ottawa, Canada. Figure 5.3 shows the relationship between road segments and road points defined by the DHS GDM. Here road paths are excluded since the relationship between a road path and its composite road segments is depicted by the name of the road segments.

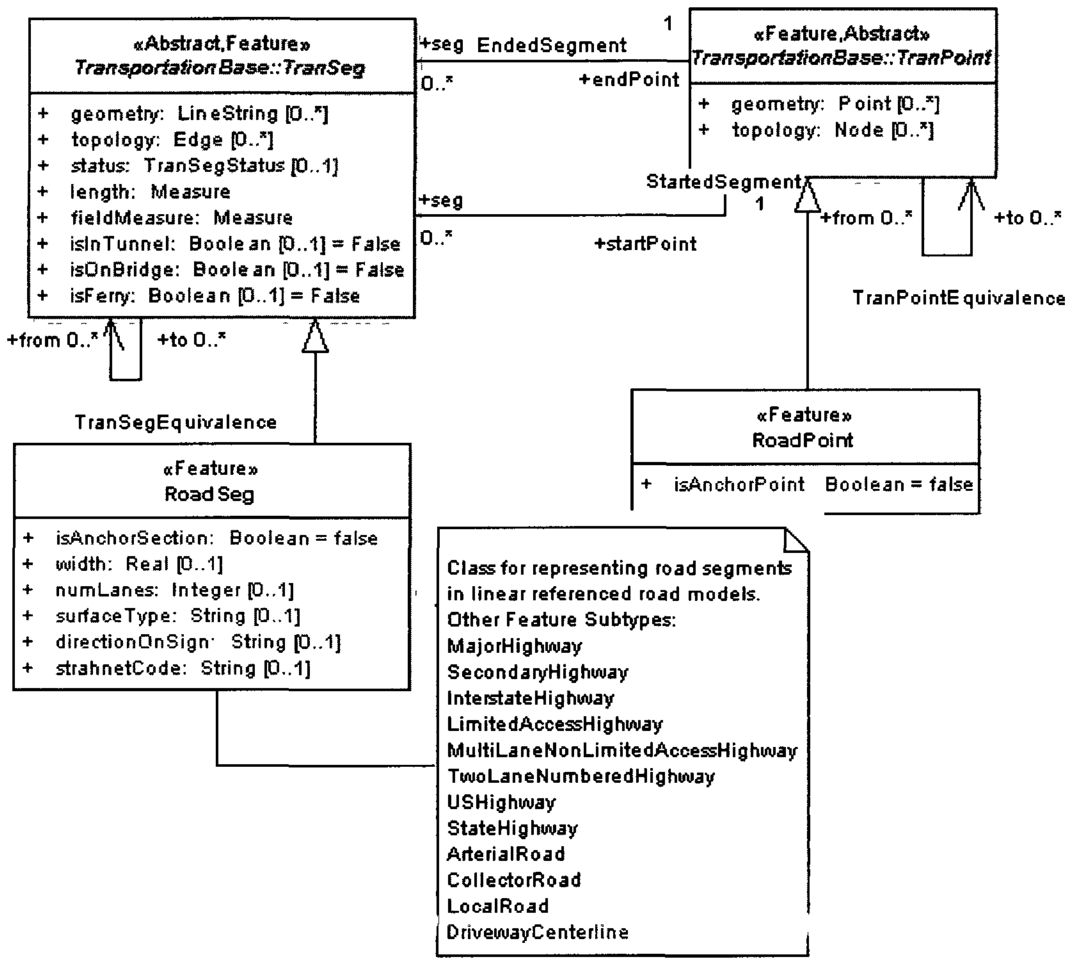

Figure 5.3: The relationship between road segments (***RoadSeg***) and road points (***RoadPoint***) and the concepts in road segments and road points. Some road classes (e.g. ***MajorHighway*** and ***SecondaryHighway***) based on different criteria are also listed in the figure (source: FGDC, 2008d).

In the Canadian road dataset, a road network is composed of road segments, ferry connection segments, block passages, road junctions, and toll points (Figure 4.2) (Geobase, 2007a). Based on the accompanying data document depicting the road feature catalogues from a segmented view (Geobase, 2007b), a road segment refers to the specific representation of a portion of a road with uniform characteristics which include road width, the number of lanes, the surface type, and the direction. This definition is similar to the Road segment concept in the domain ontology. A ferry connection segment is defined as the average route a ferryboat takes when transporting vehicles between two fixed locations on the road network. The attributes in the dataset used for depicting a ferry connection are identical to that for a road segment. Therefore, it can be deduced that a ferry connection segment is a kind of road segment in the domain ontology. A blocked passage is a point feature that indicates a physical barrier on a road built to prevent or control further access. A road junction is a point feature which bounds a road segment or a ferry connection segment. A toll point is a place where a right-of-way is charged to gain access to a motorway or a bridge.

The USA road network, however, is built up by road edges and road nodes. A node is defined as a point location representing a point in space specified by a coordinate pair (U.S. Census Bureau, 2008). Road nodes are points where changes in road characteristics occur such as road width, road direction, traffic lights, and the like. An edge refers to a visible or non-visible linear topological primitive that extends from a designated start node and continues to its end node (U.S. Census Bureau, 2008). The USA road edges includes road segments and ferry connection segments which are two different concepts in Canada road network.

Based on the above analysis, a part of the reference model for road network shown in Figure 5.4 is defined based on the road domain ontology (road model in the DHS GDM) and the application ontologies from the USA road data and Canada road data.

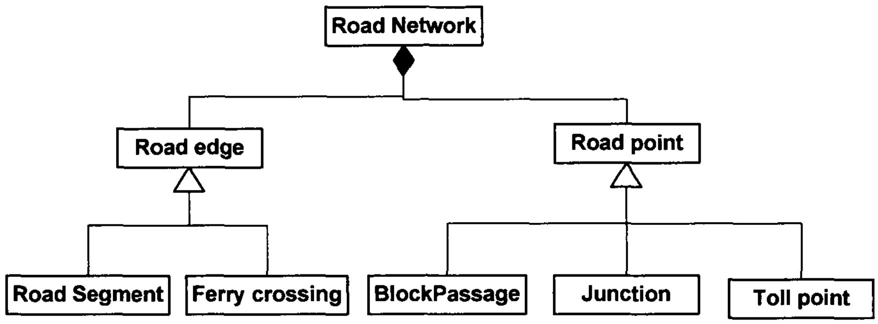

Figure 5.4: A part of the reference model which shows the road network is composite of road edges and road points. Further, road edges include road segments and ferry crossing segments, and road points contain road block passages, road junctions, and toll points.

Therefore, based on the reference model shown above and the application ontologies discussed above, the concept of a US ***road node*** refers to (Refer-to) the concept of ***road point*** in the reference model; the concept of Canadian ***road junction*** is a refer-to the concept of a ***road junction*** in the reference model and is a subclass of concept ***road point***. Hence, there is a semantic generalization relationship (S*RGeneration*) between the USA ***road node*** concept and Canadian ***road junction*** concept, so to for the US ***road node*** concept and the Canadian ***road toll point*** concept and ***block passage*** concept. The concept of USA ***road edge*** is a refer-to the concept of ***road edge*** in the reference model; the concept of Canada ***road segment*** is Refer-to the concept of ***road segment*** in the reference model. Therefore, a semantic generalization relationship (S*RGeneration*) exists between the USA ***road edge*** concept and the Canadian ***road segment*** concept. Moreover, there is a semantic aggregation relationship (SRAgregation) between the ***road edge*** concept and the ***road network*** concept in the USA (or Canada) datasets. The semantic relationships among these road network concepts from the USA and Canada datasets are shown in Figure 5.5.

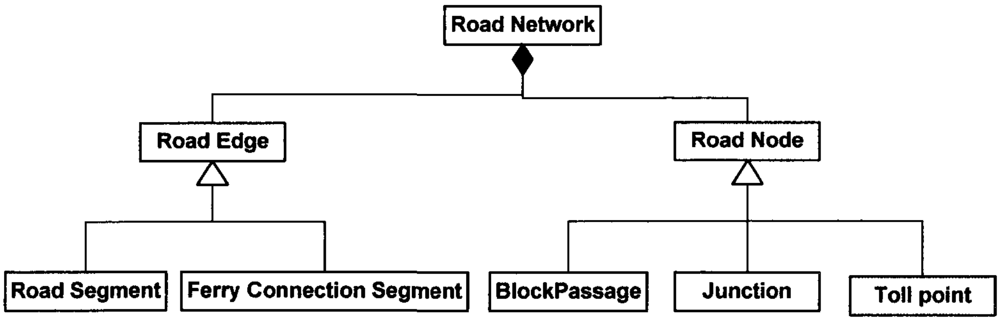

Figure 5.5: The semantic relationships among US and Canada road network concepts. Concepts with light green color background are from USA datasets; concepts with light yellow color background are from Canada datasets; concepts with white color background are for both the US datasets and Canada datasets.

Road classifications presented in both the US road datasets and Canada road datasets are also heterogeneous. In the DHS GDM, all road segments are classified based on Figure 5.3. The classification system based on road function is adopted as a part of the road domain ontology which is ***Limited Access Highway***, ***No Limited Access Highway***, ***Arterial Road***, ***Collector Road***, and ***Local Road*** (FGDC, 2008c). In the Canadian road network, road segments are classified into 13 classes which are shown in Figure 4.3 (Geobase, 2007b); while in the USA road network, road segments are classified into 13 different classes which are represented in Figure 5.6 (U.S. Census Bureau, 2008).

A lot of semantic heterogeneity is evident in the USA road classification and Canadian road classification. First, different names have the same semantic meanings. For example ***Express/Highway*** class stands for the high-speed thoroughfare with a combination of controlled access intersections at any grade; the same function road class is defined as ***Primary Road*** class in the US road classification system that refers to generally divided, limited-access highways and these roads are distinguished by the presence of interchanges. Second, the same names can also have different semantic meanings. For example, ***Ramp*** class in the USA is designated to the roads that allow controlled access from adjacent roads onto a limited access highway, often in the form of a cloverleaf interchange and these roads are un-addressable (they do not have address ranges); while in Canadian road classification system, ***Ramp*** class is designated to the interconnecting roads that provide for the controlled movement between two or more roads. A ramp in the Canadian road datasets is not only used for highway access/exit but also for arterial roads or local roads. Thus, ramps in the US road datasets are one part of the ramps in the Canada road network datasets and so form a compositional relation, *i.e.*, the ***Ramp*** concept in the Canada datasets is SRGeneration to the ***Ram***p concept in the US road datasets. Finally, the classes defined for non-arterial roads present significant differences as discussed in Chapter 4. Given the complicated semantic heterogeneity evident in the two national road classification systems, with the aid of the reference model the integration of the road classes can still be achieved shown in Figure 5.7.

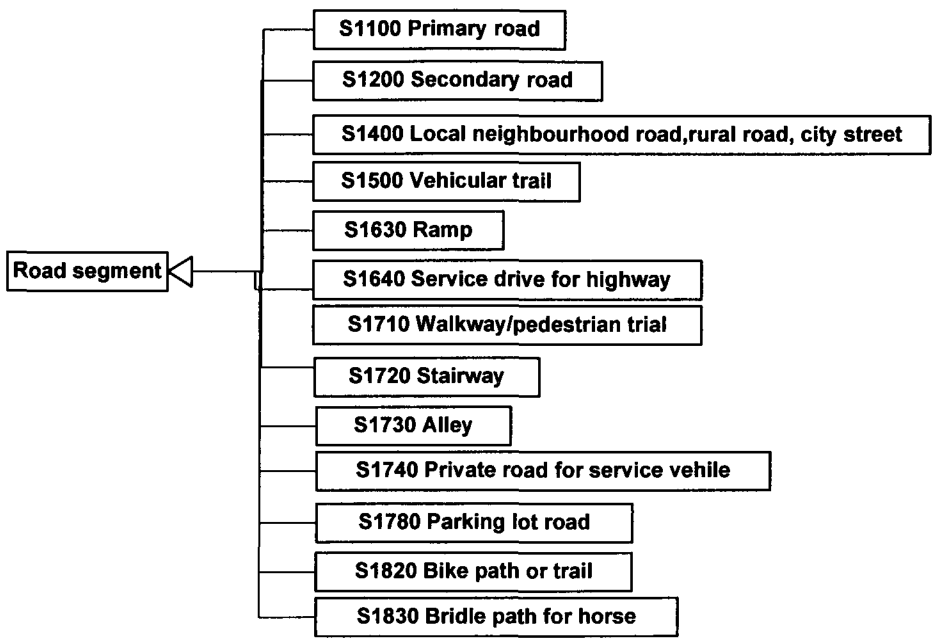

Figure 5.6: The road classification presented in the USA road network road datasets.

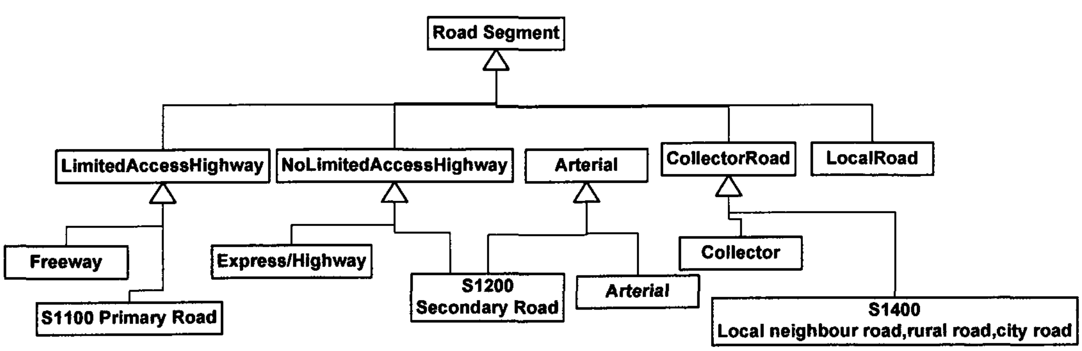

Figure 5.7: The integrated road class concepts of the US and Canada road segments. Concepts with light green color background are from US datasets; concepts with light

yellow color background are from Canadian datasets; concepts with white color background are from the reference model. The above component of ***Local Road*** is shown in Figure 4.5.

5.3 Construction of the Integrated Schemas Using MS ACCESS DBMS

The means by which one can approach the building and matching of an ontology is gaining more attention in geographic information interoperability and integration research. Some computer-understandable languages are developed to construct the ontology, for example, XML (Extensible Markup Language), GML (Geography Markup Language - Zhang *et al.*, 2003; Zhou *et al*, 2006), RDF (Aditya and Kraak, 2007), and OIL (Ontology Inference Layer) represent the major approaches. In addition, some software are designed to build and match ontologies, like protégé (http://protege.stanford.edu), COMA++ (Aumueller *et al*, 2005), and AgreementMaker (Cruz *et al*, 2007). In practice, Excel (www.microsoft.com) and a DBMS like ORACLE (www.oracle.com) or MS ACCESS (www.microsoft.com) are also employed to build and match ontologies. For example, Allen *et al.* (2008) used Excel to build up a hydro-geologic terms dictionary as the base of heterogeneous hydro-geological data integration. In this research, a MS ACCESS DBMS was applied to establish and record the integration schemas.

The Microsoft Office ACCESS DBMS provides a powerful set of tools that allows users to quickly start tracking, reporting, and sharing information (Microsoft, 2009). In addition, the ACCESS database engine called the Jet engine is the default database engine for Visual Basic for Application (VBA), which means users can directly interact with an ACCESS database using ADO (Active Data Objects) in the VBA environment without additional processes like setting an ODBC (Open Database Connectivity) data source. In this research, an ACCESS database called **frameworkdata.mdb** was created to store all the ontology-related information including:

1) the available framework data sources,
2) the integrated data schemas,
3) the conceptual and attribute ontology in integrated and application schemas,
4) the relationships among the data sources,
5) the concepts and attributes.

We provide the schema integration procedure for road segment data as an example to illustrate how the integrated schema and ontologies are built up and matched using an ACCESS database.

The start point is the table named **frameworkdata** with two fields ***Datatopicname*** and ***Haschild1*** that record (Figure 5.8):

1) Data themes have been encompassed in the framework datasets such as political_boundary, hydrography, power_line, geodetic_points, and road_network data theme, by use of the field ***Datatopicname***;
2) Whether a data theme has sub data themes, by use of the field ***Haschild1***.

The road_network data theme has five sub themes (e.g. roadsegment, roadjunction, etc.), which are marked by setting the value of field ***HasChild1*** as YES. In this case, the value of the field ***Datatopicname***, for example road_network, is linked to a table named **road_network** with one field named ***Datatopic*** which contains the name of these sub themes. However, if the value of the field ***HasChild1*** equals NO (e.g., the geodetic_points data theme), the ***Datatopicname*** field value in the table of **frameworkdata** will link to a table (i.e., the table named **geodetic_points**) containing the concepts in an integrated schema.

Again, referring to Figure 5.8, the value of the ***Datatopic*** field in table **road_network**, e.g., **Roadsegment**, directs to the table named **roadsegment** with two fields named ***filed*** and ***ontology***, which contains the integrated schema concepts and their ontology for the road segment dataset.

The information of the available data sources for the road segment data, like the shapefile name of a dataset, the country where the dataset locates, data description, and the like, is recorded in a table named **roadsegment_shapefile** with five fields which are ***Shapefilename***, ***Field,*** ***Value***, ***Country***, and ***Description***. If a sub data theme is depicted by part or whole of a dataset, it is also recorded in a table named (sub) data theme name concatenated with "_shapefile". For example in USA road segment dataset, only the records which MTFCC value equals L4165 means a ferry crossing and is depicted by setting the values of the two fields named ***Field*** and **Value** in table **Ferrycrossing_shapefile** as MTFCC and L4165, respectively (Figure 5.8).

For each integrated schema concept a table named by the (sub) data theme name and concept names (e.g., **Roadsegment_LLastHNum**, **Roadsegment_Roadfclass**) with five fields that are ***Relation***, ***Tablename***, ***Fieldname***, ***Value***, and ***Ontology*** is created to record the semantic relationship between the integrated concept and the various concepts (or sub-concepts) from the application schemas. In this table, the field named ***Relation*** records the semantic relationship between the integrated concept (which is retrieved by the table name) in the integrated schema and the concepts (which is recorded in the field of ***Fieldname***) in various data sources (which is recorded in field ***Tablename***); the ontology descriptor is recorded in the table by using the field of ***Ontology***. For example, table **Roadsegment_UniqueID** records the semantic relationship between the integrated concept ***UniqueID*** and concept ***TLID*** from the US datasets and concept NID form Canada datasets. Here is a complete example: one record in the table is (SREquvalent, Washington_roadseg, tlid, , the unique ID for a road segment in Washington state) which means that in the dataset named Washington_roadseg.shp there is one concept named ***TLID*** that is semantically equvalent (SREquvalent) to the integrated concept named ***UniqueID*** of the Roadsegment sub data theme.

In some cases, a concept in the integrated schema points to a special value of a concept in an application schema (here named subclass), like the semantic relationship among the USA and Canada road segment classification system. The USA and Canada have different classification systems for road segment classes based on road functions. The semantic relationship has been explored in Figure 5.7. This relationship is recorded via the use of a field named ***Value*** in the database. For example, a record in the table **Roadseg_roadfclass** which is (SREquvalent, Idaho_roadseg, MTFCC, has, the classification of roadsegments in the USA) means that the ***MTFCC*** concept in Idaho_roadseg dataset is semantically equivalent to the integrated concept ***RoadFClass***; moreover the field named ***Value*** has value "has", which means concept ***RoadFClass*** has some subclasses and some specified value of ***MTFCC*** are semantically equivalent to a subclass of concept ***RoadFClass***. For more detailed information of the subclasses of concept ***RoadFClass*** a table named **roadsegment_roadfclass_value** is introduced which contains the integrated road classification like ***NoLimitedAccessHighway***, ***LimitedAccessHighway***, ***ArterialRoad***, ***Ramp***, etc. For each road class, for example ***Ramp***, a table named **roadsegment_roadfclass_value_ramp** is created which contains the information like the value of field ***MTFCC*** equals S1630 from the dataset Idaho_roadseg means ramp.

Based on this recording strategy, all the concepts, attributes, and their relationships residing in the integrated schemas and various datasets are stored in the database which is also computer-readable.

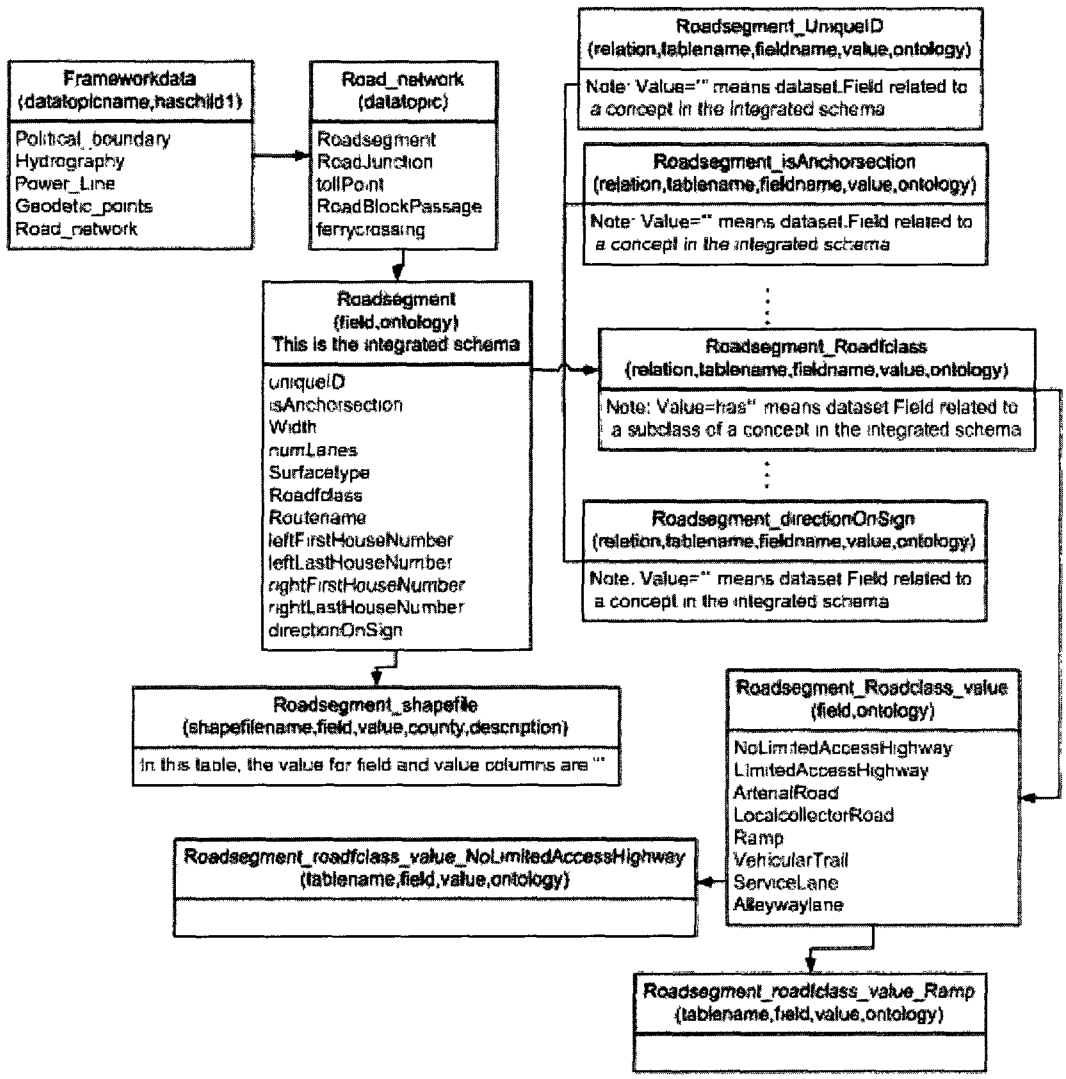

Figure 5.8: An example of how the ACCESS database frameworkdata.mdb works. The RoadNetwork framework dataset model is shown here as an example.

5.4 Prototype Interface Design and Implementation

In this research, a GUI was developed to implement seamless integration of different geospatial datasets from the border regions of Canada and the United States of America. Quite simply, the users interact with a 'front-end' on a standard workstation. The front-end is a standard windows based user interface that provides users with a choice of geospatial datasets and locations of interest. Once these parameters have been specified, the output is an integrated solution for the available heterogeneous geospatial datasets.

5.4.1 Software Used and Interface Design

The development of the interface takes advantage of the flexibility of customizing ArcGIS which allows users to add new buttons with specified utilities into the ArcMap application environment. The software, ArcMap 9.3, ArcObjects library, MS ACCESS database, and VBA were used (Figure 5.9). Two custom buttons named Data_Integration and Data_Ontology_View shown in Figure 5.10 were created and added into ArcMap application environment.

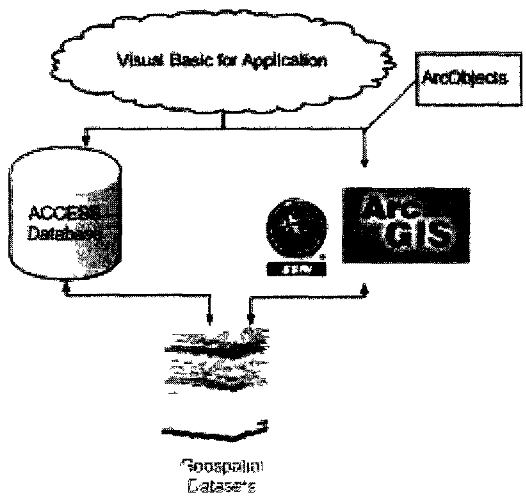

Figure 0.1: The interaction among VBA, ArcMap, ACCESS, and various geospatial data sources. The designed codes in VBA seamlessly integrate the utilities of ArcGIS by using the ArcObjects library and the ACCESS database.

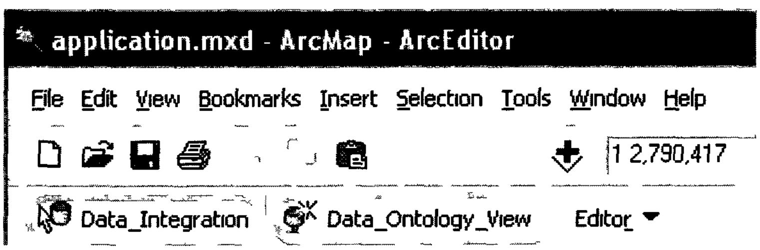

Figure 0.2: The custom buttons designed for data integration and data ontology viewing using VBA and ArcObjects were added into the ArcMap application environment.

Users click the button named Data_Integration to trigger the geospatial data integration interface (Figure 5.11). The left part of the interface lists all available cross-border geospatial datasets grouped by data themes. Users need to click the left box beside the listed (sub) data themes if they want to retrieve the (sub) data theme data. The right part of the interface provides two ways for a location selection utility (by clicking the SelectLocation button or the OpenFile button) accompanied with some common GIS tools like zoomIn, zoomOut, pan, etc. In addition, users can choose to show municipality maps, main road maps, and airport maps as contextual references to help with location selection. One way to select locations of interest is to click the SelectLocation button and then click the map. A second way to select locations of interest is by loading a custom point data file which contains all the locations of interest -users can just click the OpenFile button on the right of the interface and then specify the file as the source of locations of interest. The GO button will execute the integrated solution. When interacting with the interface, users must specify the following input parameters: the first one is named Data of Interest (DOI) which defines which framework (sub) data themes users want to retrieve; the second parameter is named Where of Interest (WOI) which specifies the centroids around which to collect the integrated framework datasets. In practice, a WOI may be a border-crossing or an emergency event location. Once the DOI and WOI parameters are specified, the user clicks the "GO" button, then an interface (Figure 5.12) pops up which allows users to input the last parameter: a buffer distance which defines the spatial coverage of the integrated framework datasets taken around the WOI centroids.

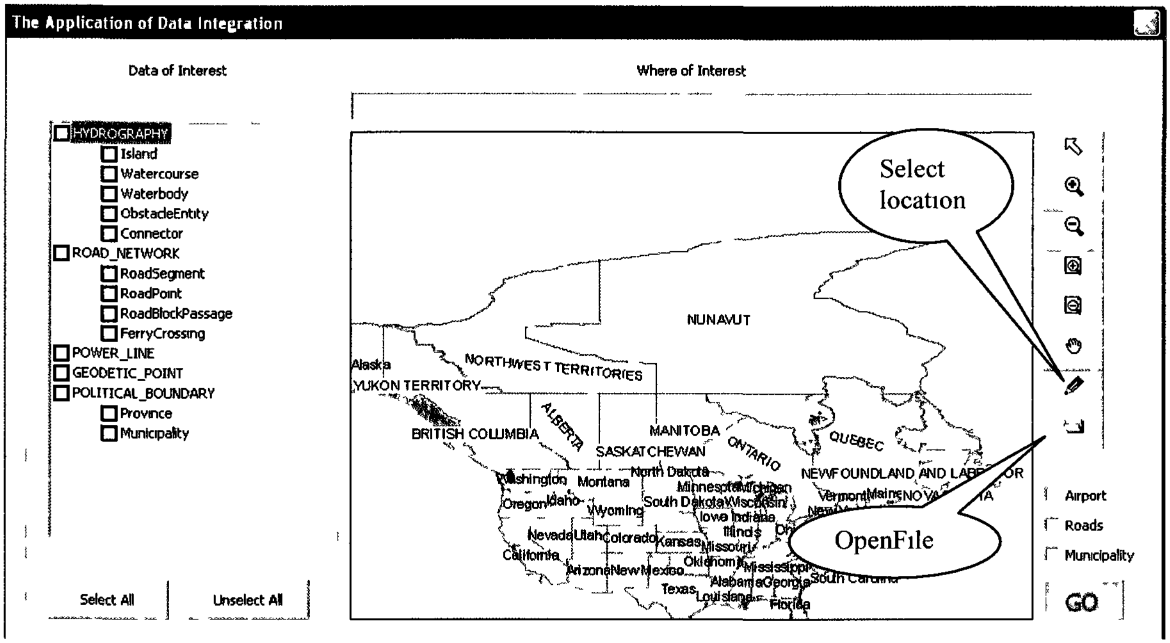

Figure 0.3: the interface for geospatial data integration specified for the US-Canada border region.

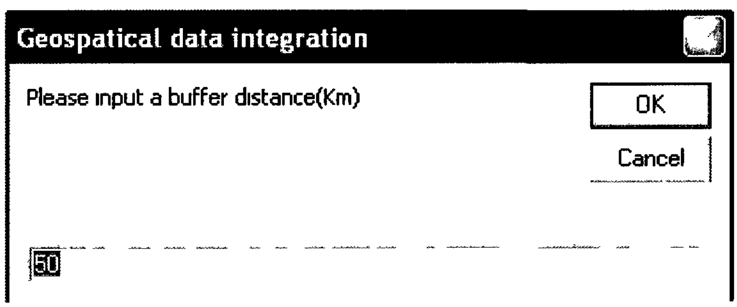

Figure 0.4: The interface designed for specifying the buffer distance.

In order to provide users with access to the ontology descriptors for the concepts in the integrated schemas and the application schemas, an interface (Figure 5.13) named Data_Ontology_Viewer is presented where users can view the semantic meanings of the concepts and their relationships residing in the reference model and various data sources. This interface is trigged by clicking the button named Data_Ontology_View which is added to the ArcMap application environment (Figure 5.10).

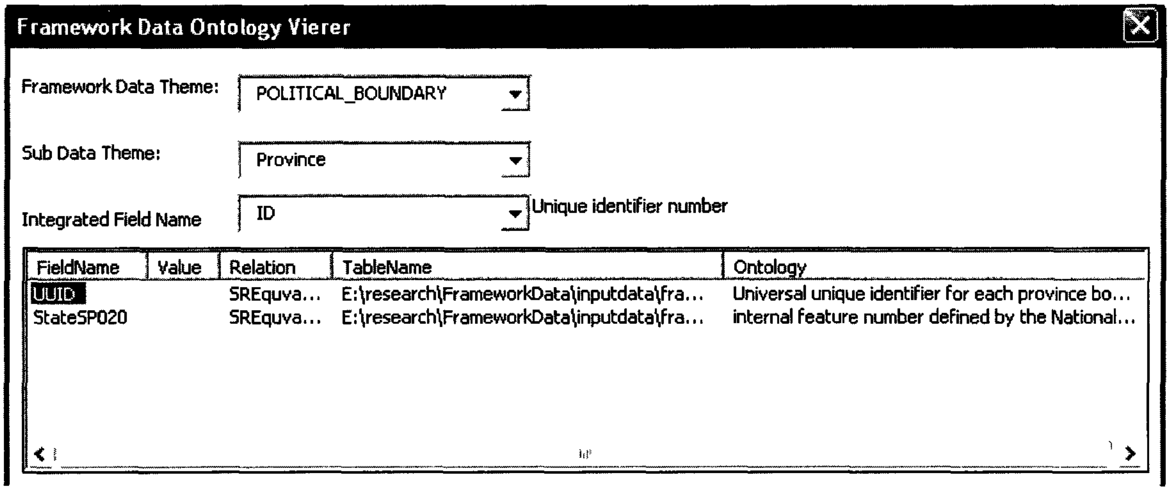

Figure 0.5: The interface designed for viewing the ontology of the concepts from the integrated schemas and application schemas.

5.4.2 The Data Flow and the Integrated Geospatial datasets

The user-driven data flow for the integration solution presents the major steps in a linear fashion (Figure 5.14):

1) Creating or specifying a point shapefile that contains all the locations of interest. Users can select locations of interest on the screen. Based on the coordinates (Latitude and Longitude) of a location selected, a point feature will be created and accordingly a point shapefile storing all the point features selected will be created. If the user has an existing point shapefile that indicates the locations of interest, it can be specified via the path and the file name in the OpenFile button in the interface;
2) Creating buffer areas. Based on the buffer distance specified by users, buffer areas will be created whose centroids are the specified points of interest;
3) Clipping the selected datasets. Once users have specified which data themes will be involved in the integration, all the datasets from the US and Canada possessing the selected data themes will be clipped by the buffer areas;
4) Revising the field names and values of the attribute tables of the clipped datasets based on the integrated schemas which are depicted in **frameworkdata.mdb**;

5) Merging the clipped and revised geospatial datasets according to the data themes or sub data themes and the geometry type of the data. The integration result is created as the way that for a selected (sub) data theme with one geometry type the result is a single shapefile and as such users only need to deal with one shapefile for a (sub) data theme holding the same geometry types without knowledge of various data sources and heterogeneous schemas and semantics. For example, if roadsegment datasets covering the USA and Canada have two geometry types which are polylines and polygons, two kinds of integration results will be created which are polyline Roadsegments and polygon Roadsegments which merge the USA and Canada Roadsegments into one shapefile, respectively.
6) Results are output to the TOC (Table of Contents) in ArcMap as well as the resulting maps and tables (an example is shown in Chapter 6 in Figure 6.2).

5.5 Summary

This chapter presented how to implement framework data integration in the North American context based on the integration methodology introduced in Chapter 4. It started with the definition of framework data; then illustrated how to integrate the data of geodetic control points, political unit boundaries, and road network. The last part of this chapter outlined how the prototype interface was designed. Chapter 6 presents a case study: how to apply the prototype and the integrated framework data to identify critical infrastructure.

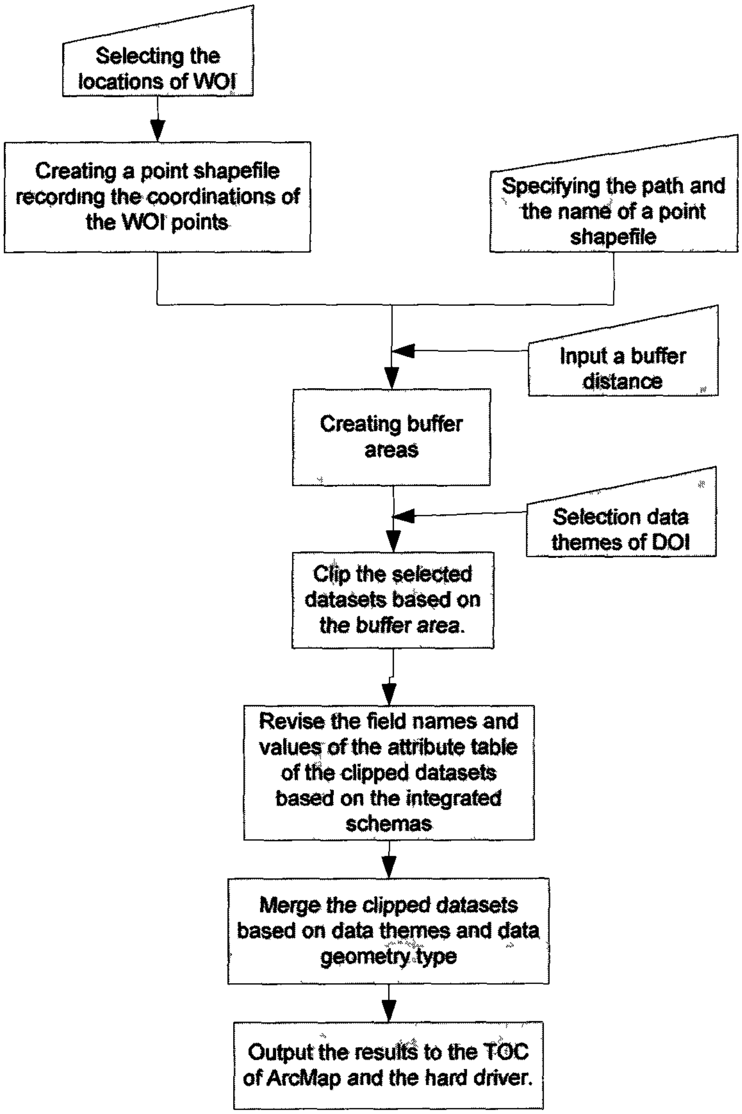

Figure 0.6: The user driven data flow chart for the integration solution.

CHAPTER 6

The Application of the Ontology-based Geospatial Data Integration---Critical Infrastructure Identification

Geospatial information interoperability and integration play an increasingly important role in management and decision-making processes that require different geo-related datasets from different organizations or data authorities, especially from different jurisdictional boundaries. For example, Riecken *et al.* (2003) presented a web-based geospatial data integration solution in the North-Rhine Westphalia cross-municipality-border environment to facilitate decision-making; Lehto *et al.* (2004) designed a prototype for a four-nation cross-border GML data service; Allen *et al.* (2008) demonstrated the use of integrated geospatial datasets for hydro-geologic information management along the British Columbia (Canada) and Washington State (US) border area; Deschamps *et al.* (2002) introduced a geospatial data integration solution for flood prediction and management in the Red River basin which crosses the Canada-US border. In the context of the Canada-US border area, more than 100 border crossings facilitate high volumes of tourism and trade. For example, about 300,000 people move across the US-Canadian border each day and in 2007 the two-way trade across the Ambassador Bridge between Detroit, Michigan and Windsor, Ontario equalled all U.S. exports to Japan (Aylesworth-Spink, 2009). Meanwhile, one of the negative aspects of convenient border crossings is, for example, more opportunity for terrorists to get easy entry points between the USA and Canada. Therefore, the longest common border in the world presents a strong need for border security and safety for the Canadian and US public and the need to plan for and mitigate man-made or natural disasters like terrorism, epidemic disease control, or critical infrastructure failures.

Recent tragedies such as Hurricane Katrina, 9/11, and the 2008 Wenchuan Earthquake in China have raised increasing attention on critical infrastructure identification and protection because of the complexity and inter-dependency of critical infrastructure between nations and regions (Petersen, 2008). Moreover, in cross-border context, critical infrastructure data are prone to be heterogeneous which results in issues with the identification and protection of critical infrastructures for disaster prevention and management (Wolthusen, 2005). Here, to illustrate the utility of this data integration research, a case study is carried out to demonstrate how integrated cross-border geospatial data can facilitate critical infrastructure identification in the US-Canada border area.

6.1 Critical Infrastructure

Critical infrastructure refers to the processes, systems, facilities, technologies, networks, assets and services that are so vital to the effective functioning of a society and economy. The loss of such infrastructure would have a debilitating impact on the security, national economic security, public safety, or any combination thereof (Public Safety Canada, 2009; Senate of The United States, 2001). Critical infrastructure involves multi-dimensional and highly complex collections of technologies, processes, and people, and therefore is inter-dependent. This interdependency causes critical infrastructure vulnerability to disasters, either natural (such as epidemic, flood, and earthquake) or human-induced (e.g., terrorism) (Public Safety Canada, 2008). Therefore, disruption of critical infrastructure could result in widespread effects, cascading across borders and organizations, rapidly escalating from local to national levels, and causing loss of life and economic damages (Public Safety Canada, 2008). For example, the Northeast Blackout of 2003 occurred throughout parts of the North-eastern and Midwestern United States and Ontario, Canada on August 14, 2003, and affected infrastructure including power generation, water supply, transportation, and communication among others. The maximum costs of this blackout totalled to US $2-10 billion (Kröger, 2008). The Wenchuan Earthquake in China left 69,197 people dead, 5 million people without housing, and exceeded US$20 billion damages. Thus, accurate critical infrastructure information is an essential resource for national preparedness to reduce risk and damages from disasters.

In fact, many countries have launched Critical Infrastructure Protection (CIP) programs. For example, USA has set up a national program of CIP in 1998 (the White House Washington, 1998), European Council has started EPCIP (The European Programme for Critical Infrastructure Protection) in 2004 (Commission of the European Communities, 2006), and Canada also has set in motion a CIP since 2008 (Public Safety Canada, 2008). To support the CIP, Canada and the USA have carried out critical infrastructure identification projects. For example, Canada launched a project name Cross-border Content and Services for Critical Infrastructure Identification aiming to help emergency-response organizations in Canada and the USA to better respond to disasters or events that jeopardize critical infrastructure (Geoconnections, 2008b); the USA Homeland Security started Homeland Security Presidential Directive 7: Critical infrastructure identification, prioritization, and protection in 2008 (Homeland Security, 2008).

6.2 Critical Infrastructure Data

Commonly identified critical infrastructures are classified into the following categories (Wikipedia, 2007):

1) Electricity generation, transmission, and distribution, like power-lines
2) Gas production, transport, and distribution like gas stations;
3) Oil and oil production, transport, and distribution like oil wells and oil pipes;
4) Communications and information technology;
5) Water supply like drinking water and water treatment plants;
6) Agriculture like food production and distribution;
7) Heating like natural gas, fuel oil;
8) Public health like hospital and ambulances;
9) Transportation systems like the network of road and railway, airports;
10) Financial services like banking;
11) Security services like military and police stations.

In the research, critical infrastructure information for the province of Ontario, Canada was collected, including:

1) Electricity facilities: pipelines and transmissions, electric power stations;
2) Gasoline service stations;
3) Water treatment plants and sewage treatment plants;
4) Public health: health care utilities;
5) Transportation systems: road network, railway network, airports;
6) Financial institutions: bank;
7) Security and safety services: police stations;
8) Education facilities: schools.

The Canadian critical infrastructure data were collected from the 2008 Enhanced Points of Interest (EPOI) data (scale: 1:50,000; coordinate system: GCS_North_American_1983) and 2008 CanMap Streetfiles (scale: 1:50,000; coordinate system: GCS_North_American_1983) produced by the DMTI Spatial Inc. of Markham Ontario.

The US critical infrastructure data collected includes:

1) Electricity facilities: pipeline;
2) Public health: health care utilities;
3) Transportation systems: road network, railway network, airports;
4) Education facilities: schools.

The US critical infrastructure data were collected from 2008 TIGER/Line census data. The scale is 1:100,000 and the coordinate system is GCS North American 1983.

In the context of the Canada-US border, the border crossings, as the main entries for the movement of goods and people, were also collected. Along the international boundary between Canada and the United States there are staffed crossings which include more than 121 road crossings, more than 35 rail crossings, more than 13 ferry crossing, and 37 unstaffed border crossings (Wikipedia, 2006).

6.3 Critical Infrastructure Identification Based on the Integrated Cross-border Geospatial Datasets.

First, one could consider numerous simplified emergency cases that would demonstrate the potential use of the integrated geospatial data in a cross-border region. Such scenarios could range from a person of interest crossing a border between the US and Canada, say, the Ambassador Bridge at the Ontario-Michigan border to the need to prepare for a possible biological agent. In such a case, critical facilities would need to be identified immediately and notified of possible security threats. One can easily imagine other scenarios such as drinking water contamination, the breach of a nuclear power plant, school or airport targets, and the like. In all these fictitious cases, an important preparedness is the identification of all the possible critical facilities like schools, hospitals, airports, power-lines, and other critical infrastructures which are within a specified travel time or distance from the border crossing, say, within a ten minute travel distance. This scenario is simplified, but it illustrates both the demand for and the potential use of an integrated geospatial data solution in a cross-border context.

The identification of the possible target critical infrastructures located within the area of a ten minutes travel distance from a border crossing requires the following steps: a) retrieve the road data around the border crossing, b) build up the road network based on the integrated road data, c) define the ten-minute-travel-distance area, and d) identify all the critical infrastructures located in the vulnerable area. These are elaborated upon in this section using the integrated solution presented in this research.

6.3.1 The retrieval of road data around the border crossing

In the hypothetic scenario, the vulnerable area is defined by ten minutes travel distance from the Ambassador Bridge crossing, therefore, the road data centered at the border crossing is the base of the identification of the possible targets. The integration prototype from the previous chapter provides the user with an easy, convenient, yet efficient way to retrieve the road data.

The WOI is the Ambassador Bridge crossing which can be selected either by clicking on the map based on the reference of main roads or by specifying the crossing shapefile. In this case study, the coordinates of all the crossings along the Canada-US border were retrieved with the aid of the list of Canada-US border crossing (Wikipedia, 2006) and the Google Maps; a layer containing all the border crossings was created and a shapefile layer designating the Ambassador Bridge crossing was created and was used to specify the WOI. The DOI here are RoadSegments and road junctions. Figure 6.1 shows that the WOI is specified by inputting the shapefile file name (e.g., the layer showing the Ambassador Bridge is called ambassadorbridgecrossing.shp) and the DOI is limited to road segments and road junctions by using the prototype. The last parameter of the prototype, the buffer distance is set as 30 km because the vulnerable area is defined as ten minutes drive distance from the Ambassador Bridge crossing, which means the speed can be estimated at a maximum of 180 km/hour. Figure 6.2 shows the retrieved road segments via the prototype after specifying the three parameters and the unified schema of the road segment attribute table. Based on the retrieved road segments crossing the border with one schema, the road network can be built up by use of the functionality within ArcCatalog.

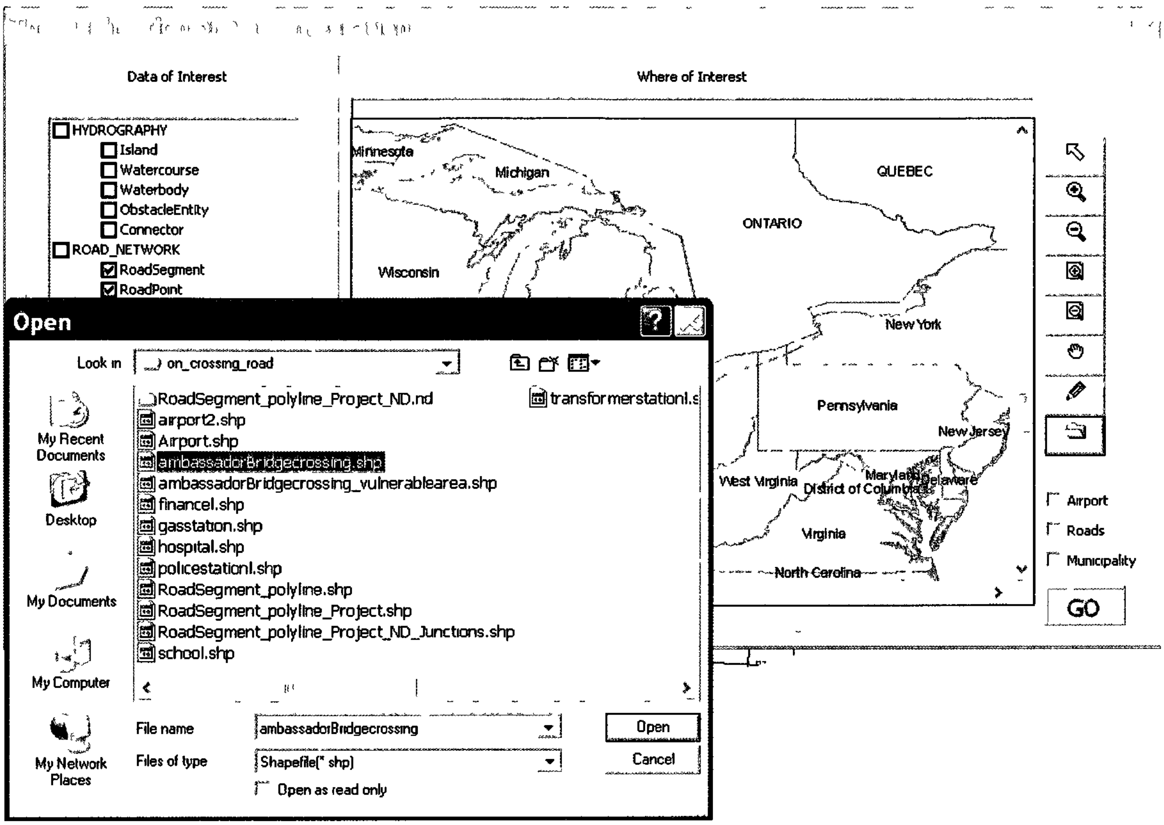

Figure 6.1: The interface of specifying the WOI and DOI in the case study

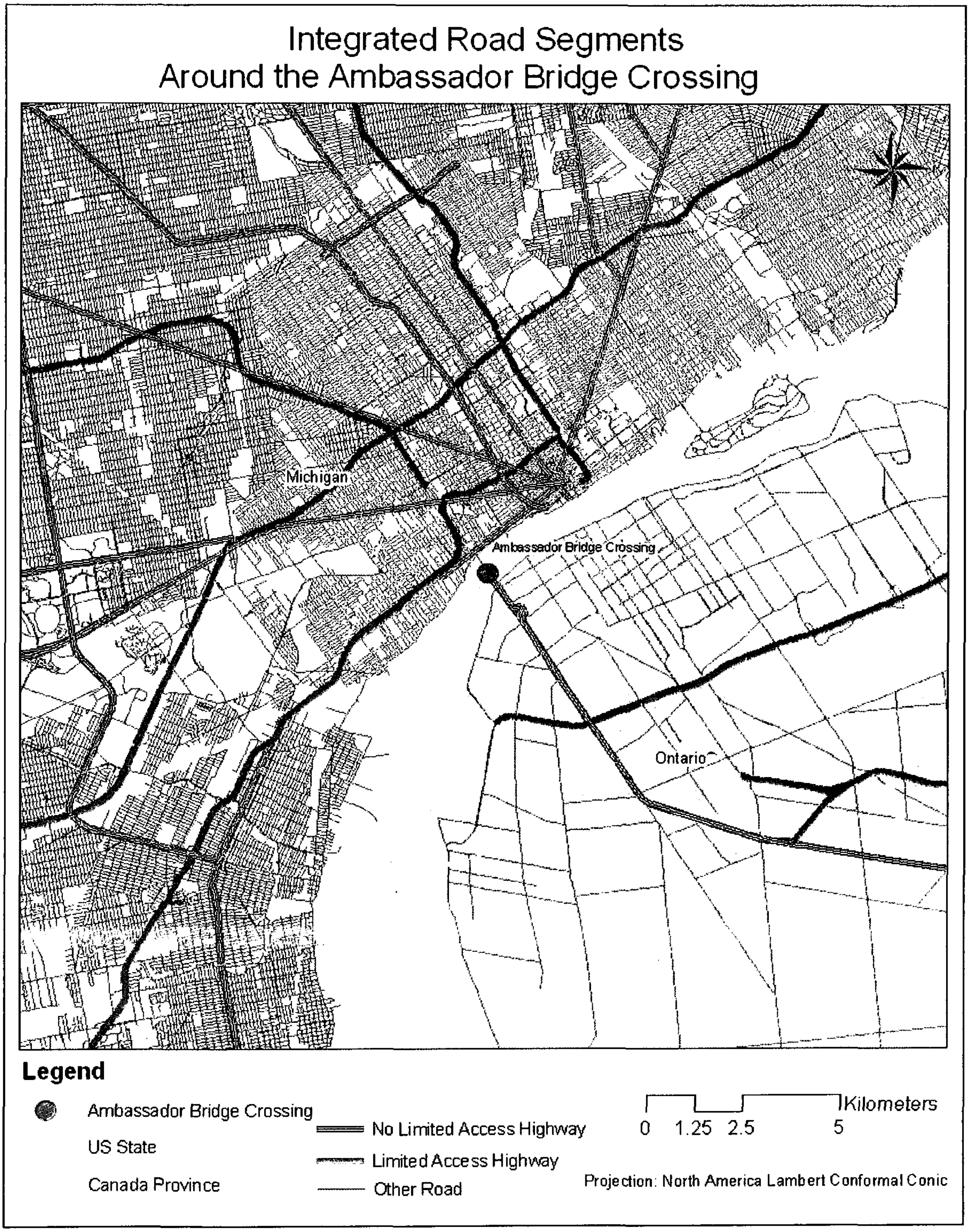

Figure 6.2: The retrieved road segment data around the Ambassador Bridge crossing.

6.3.2 The Build-up of the Road Network

The New Network Dataset utility provided by ArcCatalog is used to build the road network around the crossing. In order to get a vulnerable area determined by travel time, three pre-processes are executed that include the determination of road segment length, the limited speed and the travel time of each road segment. The length of a road segment can be calculated in ArcMap and the value stored in a new field named **Length**. In general, Table 6.1 describes the maximum speed limit on different road classes and is stored in a new field named **Speed**. Travel time for each road segment is based on a function of road segment length and speed limit (the formula: Traveltime=Length/Speed) and the result is recorded in a new field named **TravelTime**. The network dataset was modeled with single model transportation, connectivity only at end-points of road segments, global turns were utilized (which are right, left at intersections), and network cost was given as segment travel time in minutes, road hierarchy was defined by road classes. The created road network data have two datasets: road segment polylines and road junctions (shown in Figure 6.3).

Table 6.1: The maximum speed limit on different road classes.

Road Class	***Speed Limit***
LimitedAccessHighway	100
NoLimitedAccessHighway	90
Arterial Road	80
Collector Road	50
Local Road	40
Service Road	15
Other Road	0 (excluded from the road network datasets)

Figure 6.3: The created road network with road segment polylines and road junctions around the Ambassador Bridge crossing.

6.3.3 The determination of Ten-minute-travel-distance Vulnerable Area

When the road network dataset is created, the next step is to use the Network Analyst utility to determine the ten-minute travel-distance that defines the vulnerable area. The selected border crossing is designated as the input facilities; the service area is specified based on the accumulated travel time which is set to ten minutes. With the aid of the Network Analyst utility, the vulnerable area of ten-minute-travel-distance from the border crossing is specified (shown in Figure 6.4).

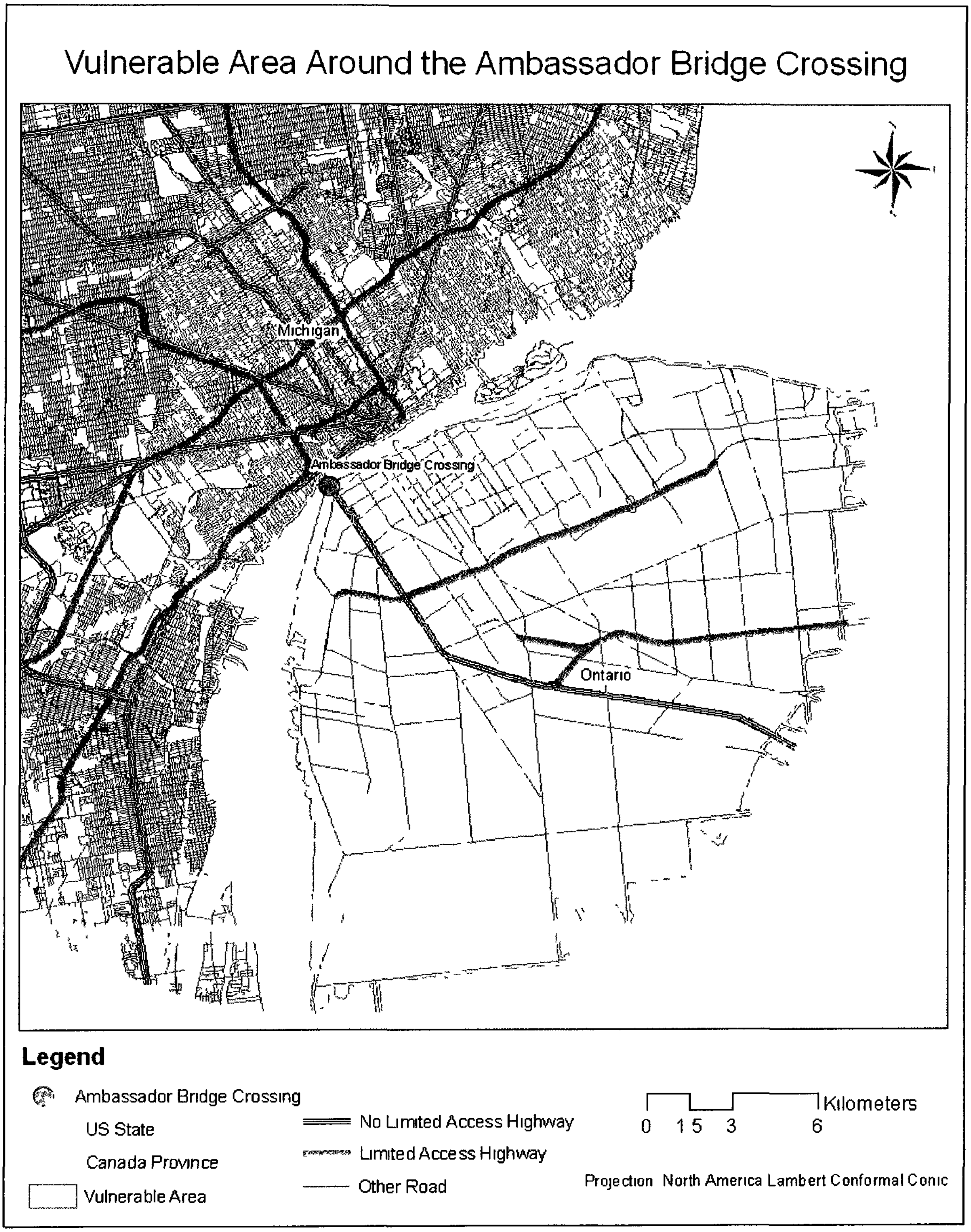

Figure 6.4: The ten-minute-travel-distance vulnerable area around the Ambassador Bridge crossing based on the build-up road network and the Network Analyst utility.

6.3.4 The Identification of All the Critical Infrastructures Located in the Determined Vulnerable Area

The Overlay Analysis utility provided by ArcMap is used to identify the critical infrastructure within the vulnerable area. Overlay Analysis is designed to address the question of "What is on top of what" (ESRI, 2007); an example of this is "which airports (or schools, roads, and etc.) are within the vulnerable area?" In this case, all the collected critical infrastructure data located within the vulnerable area is identified by overlay analysis (shown in Figure 6.5).

6.4 Summary

This chapter presented an application of the integration methodology, the prototype, and the integrated framework data: to identify critical infrastructure around the Ambassador Bridge international border crossing. Chapter 7 evaluates the methodology and prototype presented in the thesis.

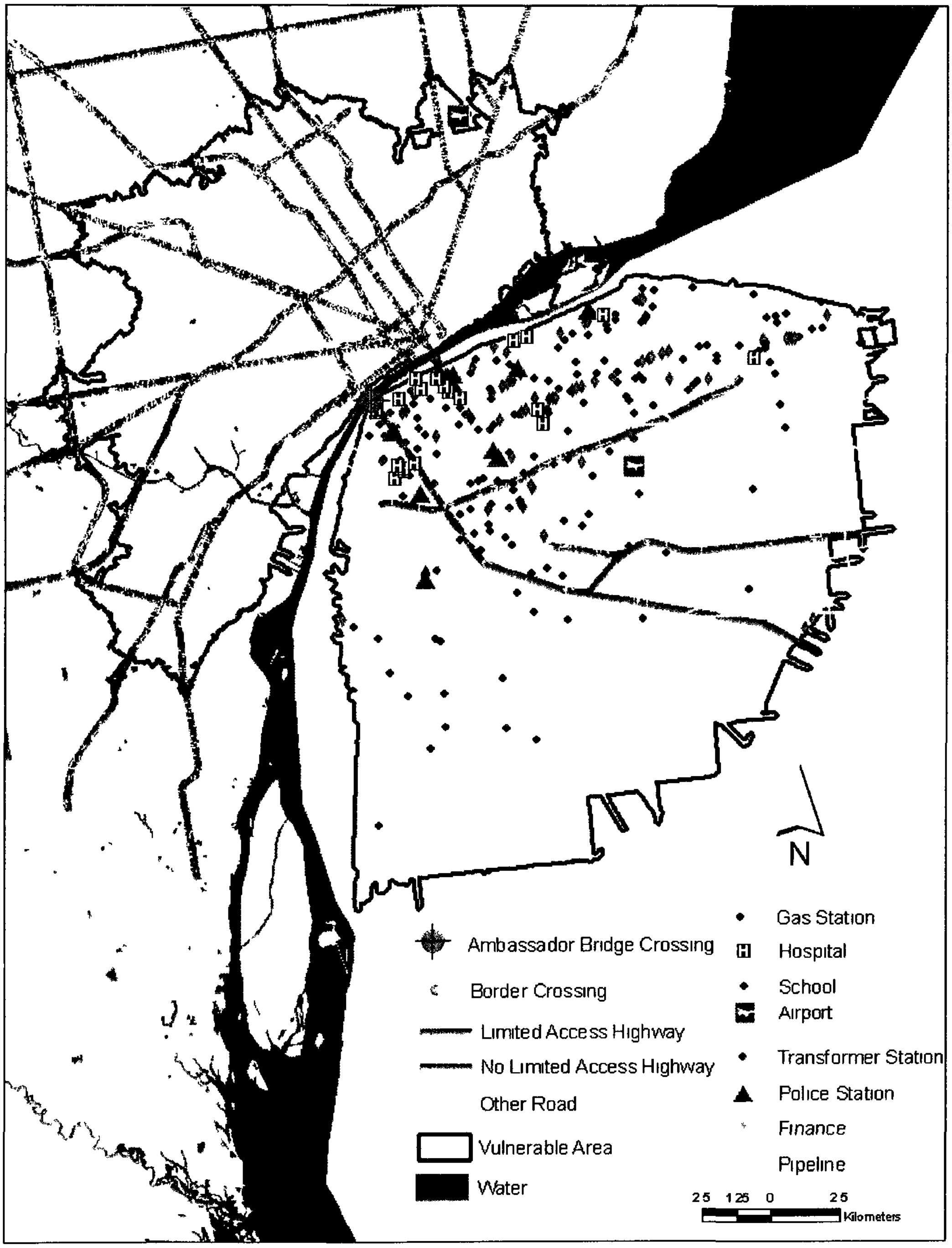

Figure 6.5: The critical infrastructure within the ten-minute-drive-time-distance vulnerable area from the Ambassador Bridge crossing.

CHAPTER 7

Discussion

The methodology presented in this research aims to integrate spatial data that has schema level semantic heterogeneity. To better understand the approach taken and its contribution to knowledge, it is pertinent here to provide context and comparison to other approaches introduced in the literature and address the fundamental questions: a) what are the advantages, disadvantages and the limitations of our approach? and; b) what were the challenges addressed and how were they overcome during the development of our approach, the prototype, and the case study? These questions are discussed in this chapter.

7.1 The Comparison between Different GDI Approaches

With the rapid development of data acquisition technologies and the awareness of the importance of geospatial information, geospatial data integration has received more and more attention over the past two decades. Many geospatial data integration approaches are found in the literature. The rough and fuzzy geospatial data integration approach introduced by Ahlqvist *et al.* (2003) can be used to integrate geospatial data with uncertainty due to vagueness and indiscernibity, and is especially useful for the integration of various classification schemes based on remote sensing data. The statistical geospatial data integration approach introduced by Walter and Fritsch (1999) is based on statistical investigations and can be applied to match objects which were captured in different data models with different geometrical representations. Safra *et al.* (2010) introduced the location-based algorithms for geospatial dataset integration, which can be applied to find corresponding objects over various datasets. The feature-based approach introduced by Samal *et al.* (2004) uses feature attributes like shape, color, texture and relations to neighboring features (i.e., the geographic context) to identify the same features with different representations in disparate data sources. All of these aforementioned approaches are designed for vertical geospatial data integration and as such these approaches are used to identify and match the same 'real world' objects represented across different geospatial datasets within the same spatial extents, thereby, allowing for successful identification of the same objects residing in those different datasets. However, because such approaches employ geometrical representation as the means for object-reality based comparison, they do not tackle the issues resulting from the semantic heterogeneity in different concept models. As such, these geometric approaches that exist in the literature do not handle or address the need for information queries and retrieval from disparate data sources that do not cover the same spatial extents for example in cross-border contexts. Handling information query on spatial objects via schema integration is necessary for successful geospatial data use for non-coincident geospatial data and our approach targets this need directly.

A second approach to integration common in the literature is the use of multi-scale geospatial databases or multi-scale GISs that can match multiple representations of the same 'real world' objects residing in different geospatial datasets (Jones, *et al.*, 1996; Devogele, *et al.*, 1996; Brisaboa, *et al.*, 2007; Balley, *et al.*, 2004). These approaches re-design the database models to establish relationships of equivalence, aggregation, and likeness between objects residing in different datasets. Semantics and ontologies are introduced in these two approaches, but they are only designed for vertical geospatial data integration. The same is true for the federated data model approach (Bererra, 1993; Devogele *et al.* 1998; Sheth & Larson, 1990) and the object-oriented modeling approach (Cartner *et al*, 2001; Abel *et al.*, 1998).

Unlike the approaches discussed above, in this research, heterogeneous cross-border geospatial datasets are horizontally integrated based on ontological integration and schema mapping. This integration approach can be used to identify and match the same real world objects represented in different geospatial datasets. Furthermore, this approach is designed to match the ontologically-similar real world objects depicted in various datasets covering different geographic areas. With the explicit semantic definitions of classes and attributes present in various data models, ontologically corresponding objects over different geographic spaces residing in different data sources can be identified. More importantly, a unified result that employs unambiguous data queries and retrieval (e.g., a shapefile) can be derived. Therefore, without the requirement for the knowledge on various data sources, users can freely manipulate heterogeneous datasets over different regions of geographic spatial coverage. As such and to reiterate, our approach allows for integration of adjacent geospatial datasets whereas existing approaches concentrate on integration of datasets covering the same spatial extents.

Ad-hoc geospatial data integration methods are a commonly used approach to integrate various data sources for a specific task, especially when retrieving data from internet sources. The advantages of ad-hoc approaches are a) they are fast, convenient and easy to use; b) there is no pre-processing required to deal with the issues resulting from data heterogeneity; c) heterogeneous geospatial data from different resources can be visualized and rendered seamlessly with the aid of some standards or specifications (e.g., Scalable Vector Graphics (SVG)) (Eisenberg, 2002; Peng and Zhang, 2004); d) users can visually interpret the data that come with geometry like distances, polygon areas, and how many cities are in a province. Therefore, an ad-hoc geospatial data integration approach is efficient and welcomed in emergency situations and applications without the need to understand data semantics.

However, *ad-hoc* integration is designed for a specific issue; it is non-generalizable, and may not be adapted to other purposes. In addition, an *ad-hoc* integration approach cannot satisfy users' requirements to query, retrieve, and manipulate data when these data present heterogeneous semantics, especially for users who are not domain experts or not familiar with the data themes, formats or metadata. For example, if users want to find municipalities in Alberta, Canada that share the border with cities in Montana, U.S., they have to explicitly state the term "city" and "municipality" in a query in order to retrieve relevant features (cities/municipalities) from both sides of the border; however, to achieve this functionality requires an implicit effort and knowledge to resolve issues of semantic heterogeneity. Unlike an *ad-hoc* approach, the integration approach designed in this research provides users with an easy way to query and manipulate various geospatial data, because after formal and explicit ontology and schema matching, the heterogeneous semantics and schemas are hidden from users and a single ontology and schema are presented to them, which make data query and exploration more convenient, simpler, and repeatable when compared to an ad-hoc integration approach.

7.2 Potential of Standardized Web Services for Spatial Data Integration

Existing web services for disseminating spatial data, like WFS (Web Feature Service), WMS (Web Map Service), and WCS (Web Coverage Service), are Open Geospatial Consortium standards that facilitate geospatial data integration and interoperability. Via the application of these standards, syntactic heterogeneity can be approached through encapsulating internal heterogeneous geospatial data structures; different map symbolizations can be harmonized by the SLD (Styled Layer Descriptor); and coordinate transformation can be implemented by WMS or WFS instances or by a WCT (Web Coordinate Transformation Service) (OGC, 2008). With these services, users can submit requests via the internet or other network and then receive responses from the web services. An OGC Web Service hides conceptual schemas of spatial data (Donaubauer, *et al.*, 2006), which means semantic interoperability or integration is still not achieved via an OGC Web Service. However, by adding schema translation and semantic translation functionality, an OGC Web Service could be used to integrate cross-border geospatial datasets disseminated on the internet (Donaubauer, *et al.*, 2006; Jaenicke, 2003). Taking advantage of the OGC Web Services, to integrate different data sources on the internet would be an efficient way to convert the prototype designed in this research into a generalizable network application.

7.3 Evaluation of the Ontology-base Geospatial Data Integration Methodology

Many researchers have applied ontologies to data integration procedures for different purposes. There are a number of surveys that evaluate and compare the proposals or approaches that focus on information integration that employ ontology (Kalfoglou and Schorlemmer, 2003; Euzenat and Shvaiko, 2007, Wache *et al.*, 2001; Klein, 2001; Buccella *et al.*, 2009). Ontological criteria proposed by Buccella *et al.* (2009) (Figure 7.1) is particularly useful for evaluation our current integration methodology and the prototype.

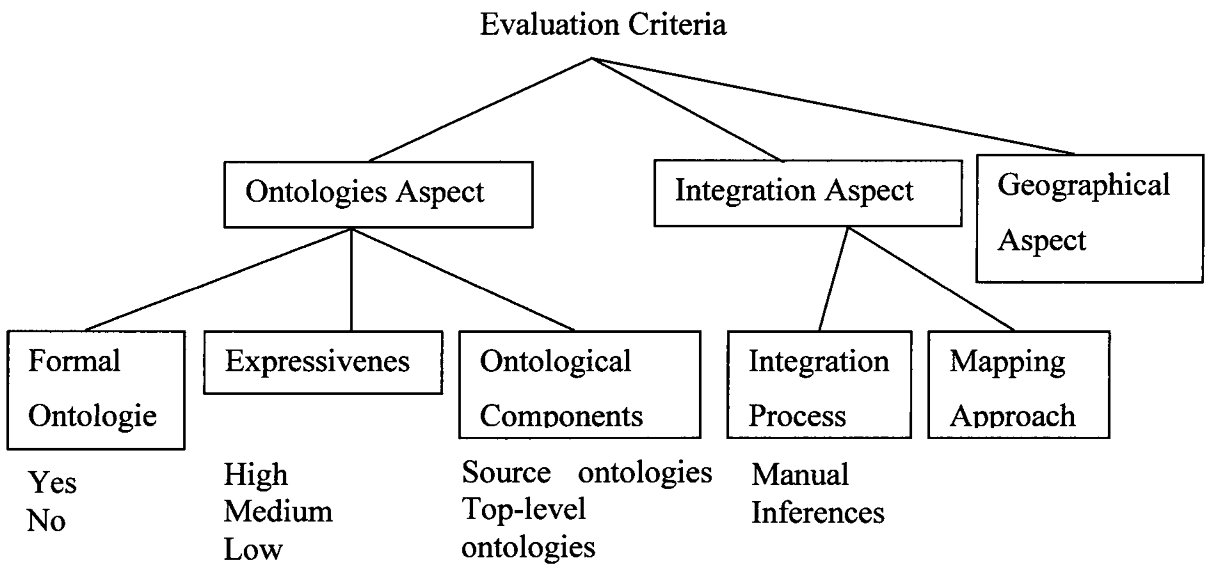

Figure 7.1: Evaluation criteria for data integration proposed by Buccella *et al.* (2009)

The criteria shown in Figure 7.1 contain three aspects: ontologies, integration, and geographical. The ontologies aspect of Figure 7.1 evaluates if ontologies are formal or not, how efficient ontologies can depict the feature semantics, and which levels of ontologies are applied. Integration evaluates if the ontology inference process is carried out manually or automatically and the defined schema mapping approach. Geographical evaluates if the integrated content is graphical or not. An ideal process to integrate geographical datasets is the one with automatic infererence of semantic relationships by using formal and high expressive ontologies. The evaluation of the integration approach introduced in this paper based on these criteria can uncover the limitations or advantages of our approach and also point out future work directions.

7.3.1 Formal Ontologies Criterion

On the ontologies of the evaluation criteria, Buccella *et al.* (2009) suggested that only those ontologies that are represented by a logical (or formal) language are formal. For example ontologies represented by Web Ontology Language (OWL) or Resource Description Framework (RDF) are formal ontologies (*c.f.* Buccella *et al.*, 2010). The advantage of using formal ontologies is that semantic reasoners (reasoning engine, rules engine), a piece of software able to infer logical consequences from a set of asserted facts or axioms (c.f. Jena described in McBride, 2002 and Pellet described by Sirin, *et al.*, 2007), can be applied to infer implicit relationships between concepts and attributes. On the contrary, informal ontologies are those represented by such tools as a controlled vocabulary, a thesaurus, informal taxonomies, and others (De Bruijn, 2003) like WordNet used in our mapping application.

The ontologies developed in this research (i.e. application ontologies, domain ontologies, and the reference model) are informal ontologies. Instead of using a formal language to represent ontologies, a DBMS is applied to record ontologies with a controlled vocabulary from WordNet. The relationships between concepts and attributes are depicted by the relationships between the tables and the fields in the designed ontology database. For example, the semantic aggregation relationship between the concept of ***road network*** and ***road segments***, ***road points***, and ***toll points*** are depicted by a table called **Frameworkdata** and a table called **Road_network**. The benefit of using a database is that the ontologies developed in this paper are efficient enough to be processed by a computer and more importantly, they are easy to incorporate into ArcMap via the ArcObjects library. However, one of the limitations of these informal ontologies is that users cannot utilize existing reasoners to infer semantic relations residing in the ontologies. Thus the ontologies developed in this paper do not have enough expressive power to represent the rich semantics of the real world, compared to formal language ontologies where relationships like 'Connects to', 'Is On', or 'Under' can be conveniently defined using a natural language (Varanka, 2008). In order to infer the explicit relationships from the ontologies developed in this paper, designers need to navigate the tables with the help of documents. Thus, a future step for this research would involve the development of a new GUI that facilitates the users' understanding of the relationships between the concepts and attributes residing in various datasets.

7.3.2 Ontology Expressiveness Criterion

On the ontologies aspect of the evaluation criteria, Buccella *et al.* (2009) defined three levels of ontology expressiveness. A low level of ontology expressiveness implies that the language used has limited capabilities to represent properties and axioms of ontology definition. For example, the ontologies proposed by Hakimpour (2003) have low expressiveness since only the generalization/specialization relation is modeled. A medium level of ontology expressiveness refers to the ontologies that include the generalization/specialization relation and also other relations like 'whole-of' or 'part-of' relations (*c.f.* Kavouras *et al.*, 2003). A high level of ontology expressiveness means non-restricted representation of ontologies (*c.f.* Wang, *et al.*, 2007; Varanka, 2008; Sotnykova *et al.*, 2005; Buccella *et al.*, 2010). In order to develop ontologies with a high level of expressiveness, some researches proposed ontology-driven GIS (ODGIS) where geospatial relationships, topological relationships, and conceptual semantic relationships are depicted by defined ontologies (Fonseca, 2002; Fonseca and Egenhofer, 1999; Fonseca, *et al.*, 2002).

In this research, three relations are defined which are SREquivalent (semantic equivalent), SRGeneration (semantic generalization), and SRAggregation (semantic aggregation), which equate to a medium level of ontological expressiveness. However, these relations do not fully depict geospatial relations such as 'including', 'neighboring', and 'like'. For example, our ontology cannot address questions such as: Which provinces or states are neighboring Ontario? However, the question can be addressed by using the Spatial Analysis model of ArcMap, which is one of the benefits of adopting ArcMap as the application platform for integration – we leverage rather than duplicate existing GIS functionality.

7.3.3 Ontological Component Criterion

The last part on the ontologies aspect of the evaluation criteria is ontological components. An ontological component defines the level of ontology (i.e. application level, domain level, or global level) developed in an application. For example, some applications may only build application ontologies, while some may also develop domain and global ontologies. Such information is useful for deducing which integration process, either hyper ontology or single ontology approaches, can be applied.

In this paper, a reference model, domain ontologies, and application ontologies are developed to facilitate data integration. The domain ontologies are developed based on the standards and specifications shared by Canadian and US GICs, which are the base of the development of the reference model. The reference model in this paper is crucial in the cross-border geospatial data integration process. All the application ontologies are mapped to the reference model through the three defined ontology relations (i.e. SREquivalent, SRGeneration, and SRAggregation), which is a type of hyper ontology mapping approach. This approach avoids conflicts resulting from independence of domain ontologies. Moreover, this approach provides users and designers with a free ability to add new data sources to the integrated datasets.

7.3.4 Integration Process Criterion

On the integration aspect of the evaluation criteria, Buccella *et al.*, (2009) suggests evaluating whether a process implements a manual, semi-, or completely automatic integration process. A manual process involves no methods or tools used to create an ontology mapping and users or experts have to examine application ontologies to explore concepts and attribute relations. Manual mapping is time-consuming work and requires expertise in the subject/domain of the application. Moreover, manual mapping is prone to error when dealing with many hierarchies of concept in a series of heterogeneous datasets.

On the contrary, some methods or tools can be applied to map ontologies in semi- or completely automatic processes, like using reasoners to infer concept similarity. Researchers have developed some methods or algorithms to measure or evaluate concept similarity (Rodriguez *et al.*, 1999; Rodrigues and Egenhofer, 2004; Schwering and Raubal, 2005; Zhou *et al.*, 2006; Maedche and Staab, 2002; Kavouras, *et al.*, 2005). Semi- or completely automatic ontology mapping using well-designed methods or tools is a resource- and time-saving process and the result is more accurate as long as the methods or tools are well designed. Another more important advantage of semi- or completely automatic ontology mapping is that the mapping process is easy to repeat when new data sources are added.

In this research the mapping between the reference model and the application ontologies is achieved manually. In order to facilitate adding new data sources, an interface needs to be designed in the future. This future design will provide users with abilities to add data sources, to input data schemas, to browse existing ontologies, and then to map new ontologies to the existing ones by dragging or clicking. Through such an interface, the ontology mapping process would be considered as a semi-automatic process.

7.4 Some Challenges

The prototype software application designed in this research provides users with abilities to easily access (i.e., query and retrieve) different data sources. Some challenges occurred in the process of designing the prototype. Two of these challenges include how to process data with geographic discrepancies and how to deal with map scales that contain different levels of generalization. The following will discuss these two challenges.

7.4.1 Geographic Discrepancies

Geographic discrepancies include feature density variations (Figure 7.2), feature discontinuities (Figure 7.3), attribute discontinuities, and horizontal offsets (Figure 7.4)

(Deschamps *et al.*, 2002). Feature density variations may be caused by different levels of generalization (i.e. different mapping scales), different user perspectives, or different data purposes. Feature discontinuities occur when features appear in one dataset but not in another dataset and may be result from different user perspectives, different data scales, or just mapping errors present in datasets. Attribute discontinuities occur when some attributes are present in dataset A while not in dataset B.

Attribute discontinuities are clearly evident in this research. For example, the attribute of road surface type is included in the Canadian road network but not in the USA road dataset. A solution to this issue is to construct a reference model based on universally (in this research, Canada and the USA) accepted standards and specifications and then to set the values of the missed attributes to NULL so that users can still pose queries relating to these missed attributes. The results, however, would be a NULL or empty set. This solution does, however, provide the opportunity to add the missed attributes once they are available.

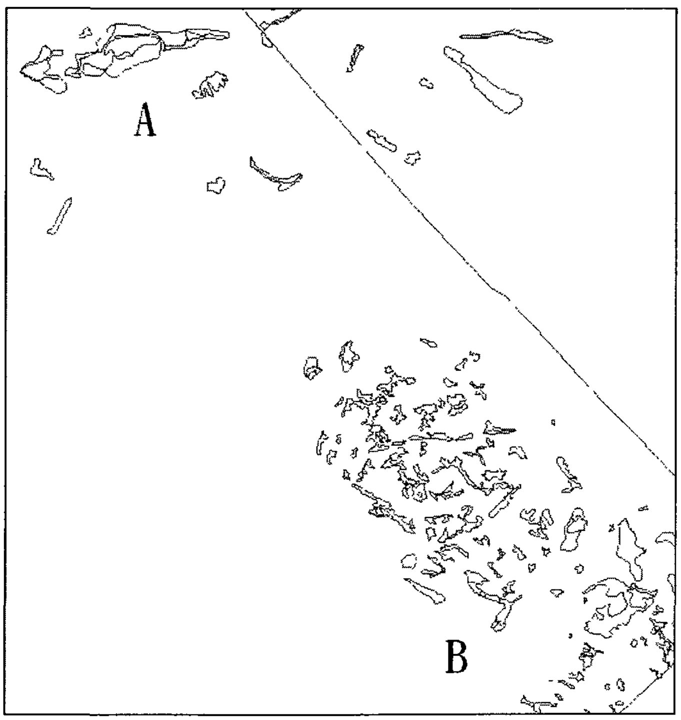

Figure 7.2: An example of feature density variations. The features in dataset B are denser than dataset A.

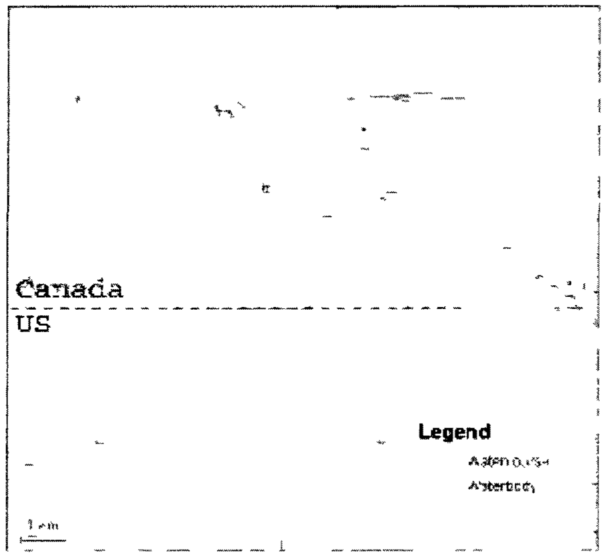

Figure 7.3:Wetland discontinuity in datasets at the border of Canada and U.S. (Deschamps et al., 2002).

Horizontal offsets occur when adjacent (bordering) geometry features are present in both datasets, but they are offset horizontally to some extent. In this research, horizontal offsets (Figure 7.4) occur frequently. For features in Figure 7.4, after checking the road attributes (like road name and the attributes of road nodes) and Google maps, road segment 1 and 2 across the border are connected in reality, so are road segment 3 and 4. However, these road segments in the two datasets integrated in this research are not connected and have some horizontal offsets. In practice, users can snap the road nodes together using standard topological editing tools in a GIS. For instance, in this case, users can snap the node of Road 2 to the node of Road 1, and snap the node of Road 4 to Road 3. This solution will make building up topological relationships between the road segments in these two different datasets possible; however, this will introduce some horizontal uncertainty. It may change, for example, the relative position between road segments 2 and 4, and the distance between node A and B. Another possible solution to horizontal offsets is to adjust the offsets based on higher resolution remotely sensed images instead of snapping nodes. The first solution was applied in this research; however, how much uncertainty is introduced needs to be addressed in the future work.

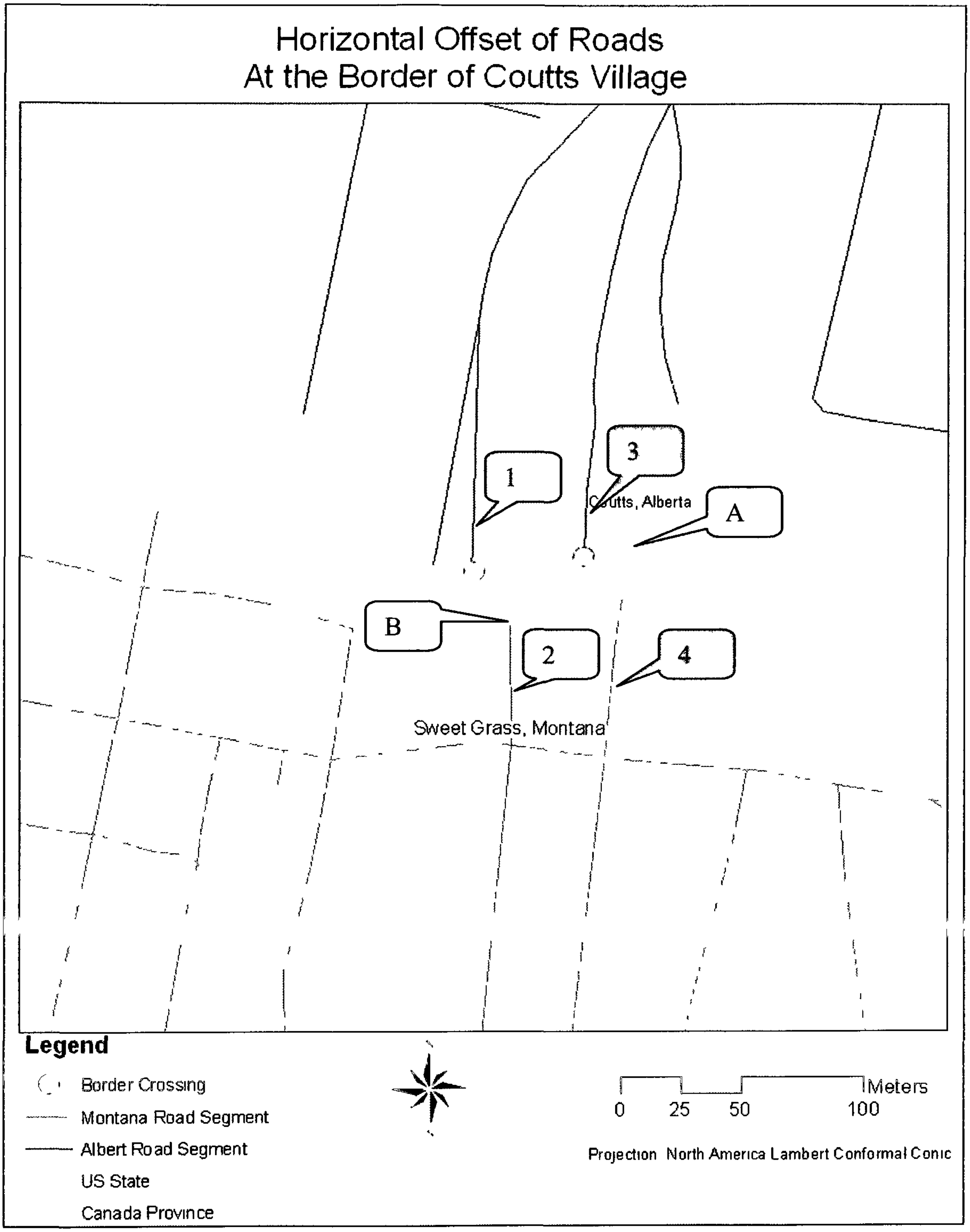

Figure 7.4: Horizontal offset of the roads at the border of Coutts village, Alberta and Sweet Grass, Toole, Montana. In reality, Road 1 and Road 2 are connected, and so are Road 3 and Road 4.

7.4.2 Scale Compatibility Issue

Spatial analysis models are scale-specific or scale dependent, which means for instance models developed for large scale maps need to be adjusted when used in small scale maps or integration situations (Beven, 1995; Lilburne, *et al.*, 2004; Heuvelink, 1998; Lilburne, *et al.*, 2003). The results derived from scale incompatible datasets are misleading and even invalid in many cases (Lilburne, *et al.*, 2003). Therefore, a specific model should be provided for each scale in practice. The reasons are as follows: (1) the processes for different scales are different, so models used at different scales should focus on different processes; (2) the data availability at different scales is different, which sometimes means that the model parameters cannot be derived; and (3) the model input and model output in different scales have different sizes, shapes, and orientation of the entities (Heuvelink, 1998).

Therefore, although semantically similar features in different datasets can be identified, applying the same model to the integrated multi-scale datasets will inevitably introduce uncertainty. For example, a road network along a border is recorded in dataset A and dataset B. Dataset A shows the main road central lines and road junctions, while dataset B is more detailed which shows not only main road central lines and road junctions but also local roads and ramps. One of the problems that become clear from an analytical perspective is how can road network analyses such as shortest route selection be implemented based on these two datasets with different scales. One approach is to adopt the model that is suitable for dataset A as the model used for the integrated dataset on the condition that the output meets a predefined quality standard. This approach can be implemented if the scales of the datasets do not vary too much. The second approach is to adopt different models for various dataset with different scales, which may be difficult to implement in practice. The third approach is to scale up a finer scale dataset (like dataset B in Figure 7.2) to a courser scale dataset (like dataset A in Figure 7.2). Scaling up datasets (i.e. map generalization, database generalization, or cartographic generalization) has been an intensively discussed topic in the cartography literature (Mustiere, 2005). Some GIS software also provide dataset up-scaling functionality like

ArcGIS (www.esri.com) and MapInfo (http://www.pbinsight.com). The issue of scale compatibility is a part of the future work and was not considered in this research.

7.4.3 Temporal Discrepancies

The currentness of geospatial datasets can induce geographic discrepancies as well as ontological and semantic heterogeneity. The temporality of geospatial data and issues induced therein for a given integration strategy need to be addressed. These issues are a topic for further research.

7.5 Summary

This chapter compared the main approaches for geospatial data integration, especially the ad-hoc approach. We suggest that the integration of geospatial data based on the OGC Web Services and ontology matching will make the prototype software application designed in this research more practical. The methodology and the prototype presented in this paper were also evaluated by the criteria introduced by Buccella *et al.* (2009) from the aspects of ontology and the integration process, which provided a means to specify the limitations of our methodology and prototype. The last part of this chapter discussed three main challenges when integrating geospatial data across the border and further applying these data in spatial analysis.

The next chapter of this research summarizes the work on the ontology-based geospatial data integration and the future directions of this research are also discussed.

CHAPTER 8

Conclusions and Future Work

This research focused on developing an ontology-based methodology to integrate heterogeneous geographic data in cross-border contexts. The methodology provided for the integration of geospatial information from the conceptual world and is based on semantic contents of geospatial data rather than based on data formats, geometric representations, or feature locations. Based on the methodology, a prototype with a GUI was designed to illustrate integration in a North American context given the semantic and schema heterogeneities present in international datasets. Moreover, a case study aiming to identify the critical infrastructure around the Ambassador Bridge crossing was carried out using the prototype in order to illustrate the kinds of questions that can be addressed when adjacent datasets are integrated.

8.1 Summary of the Thesis

This paper investigated the use of ontologies as the foundation for integrating heterogeneous geospatial data. Ontology integration is an essential step to achieve schema matching and consequently data integration. Ontologies were chosen as the basis of this methodology because ontologies can represent real world objects by using a collection of unambiguously defined and structured concepts which include classes, attributes, constraints, and relationships. Furthermore, ontologies capture the semantics of the data, which can be used to establish the basis for information sharing and reuse inside or among GICs. In addition, ontologies can be represented in a formal language and therefore a reasoner can be applied to simplify the semantic relationship inference.

A conceptual framework for ontology-based geospatial data integration was introduced. To find a unifying concept of space or to find a common ontology is difficult, so it is necessary to deal with multiple perspectives of the geographic world. Starting from real world objects, different GICs conceptualize the objects from different views and therefore result in heterogeneous semantics (i.e., different concepts, attributes, and relationships, etc.). The refined reference model presented in this methodology acts as a bridge to map or to match various concepts and attributes from different GICs. The introduced methodology and the integration process provide a practical approach to integrate various geospatial data whether or not those data share a common spatial extent or are adjacent in space. Finally, the methodology deals with the integration of geospatial data at different levels of detail by matching ontologies.

The hybrid ontological approach is applied to map schemas to solve the problems resulting from different views of the real world. In this research, the DHS GDM was selected as the domain ontology; however, instead of simply adopting the DHS GDM to implement the integration procedure, a reference model was developed that acts as a global ontology to achieve schema mapping and therefore data integration. The reference model was developed not only based on the DHS GDM but also extending that to the concepts in the application ontologies, which eases and simplifies the communication and the matching of heterogeneous application ontologies and schemas. More importantly, the hybrid ontological approach together with the reference model allows users to add new geospatial data to the integrated database.

The selection of a domain ontology as the basis to construct the reference ontology is important in the context of crossing an international border. The specifications and standards used in each nation to produce and manage geospatial data are apt to be different due to the differing politics, cultures, history and technologies in each region and as such geospatial data are more likely to be heterogeneous. For example, to integrate geospatial data produced by Canada and China respectively will be a large challenge since there are differences that are even more substantial because of the distance between the political and social cultures. Fortunately, in the context of the Canada-U.S. border, geospatial data from Canada observe the CGDI and the data from the USA follow the DHS GDM. The DHS GDM stems from the standards and specifications endorsed by the FGDC. The CGDI and the standards and specifications of the FGDC are compatible and have a high degree of alignment since both of them adopt the specifications and standards of the OGC and ISO. Therefore, choosing the DHS GDM as the domain ontology eased the integration procedures.

In this research, ArcMap was chosen as the platform to carry out the integration. The advantage of this is that the prototype can leverage the powerful functionalities of geodatabase management and spatial analysis provided with ArcMap. However, one challenge in using ArcMap would be how to incorporate a reasoner to infer the ontology relationships with a formal ontology language.

The case study was presented to illustrate the utility of the integration approach for cross-border security. The case study identified the critical infrastructure around the Ambassador Bridge crossing, and answered the research questions proposed in the first chapter which were: 1) how can geospatial data integration be undertaken effectively in the North American context given the semantic and schema heterogeneity presented in diverse international datasets? And; 2) can such an integration procedure be streamlined via a software solution? To use the prototype is straightforward: users only need to set the locations (points, polylines, or polygons), and then the prototype will present the required geospatial data with unified schemas and ontologies. The processes of ontology matching and schema mapping and the retrieval of the base layers and critical infrastructure data are hidden from users. Moreover, there is no requirement for users to understand metadata, and users will not notice that query information results come from two or more heterogeneous data sources.

The methodology and the prototype can not only be used to integrate geospatial data with overlap, but more importantly can be used to integrate geospatial data without overlap (bordering or not). The application of the methodology and the prototype is not exclusive to critical infrastructure identification. It can be used to geospatial data upgrade, environment management and protection, disaster management, and others.

8.2 Future Work

There are a variety of issues that remain to be resolved. Firstly there are methodological issues that require elaboration such as applying a formal ontology language and reasoner, integrating geospatial data on the internet, and uncertainty analysis for the application of the integrated geospatial data. Secondly, practical issues are present in the research that regard how to upgrade the prototype, how to add a new interface or how to integrate remotely sensed data.

The literature introducing geospatial data integration methodologies applies a formal ontology language to record ontologies and a reasoner to investigate the inferred relationships between classes and attributes. For example, Protégé uses IDL or OWL and Jena (a kind of Java) to build ontologies and the reasoner. It would be beneficial for the prototype application to incorporate a formal ontology language and a reasoner to record and infer the ontologies in the future.

In reality, more and more geospatial data are disseminated on the internet because of the continuing development of information technologies and WebGIS. Users can freely download and retrieve geospatial data from websites. Thus, there is an increasing demand for instant integration of geospatial data disseminated on the internet. Our prototype is a stand-alone application with the geospatial data residing on local hard drives. Therefore, in order to make the prototype more useful it would be necessary to add functionalities that allow users to browse internet for geospatial data and to integrate the required data and then present users with an integrated and unified result.

Uncertainty is considered as an integral part of geospatial data resulting from data acquisition, data format transformation, and spatial analysis (Chrisman, 1991). Therefore, uncertainty analysis is an important step in the application of geospatial data to any specific purpose. The reasons for considering uncertainty are that 1) uncertainty residing in raw geospatial data will contaminate the application result; 2) uncertainty can be propagated through GIS operations like spatial analysis, and more importantly, the propagated uncertainty may be much larger than the uncertainty in the raw data (Crosetto & Tarantola, 2001); and 3) uncertainty analysis can provide users with a reliability or confidence level as for how much the data or results can be trusted (Crosetto & Tarantola, 2001; Heuvelink, 1999; Heuvelink & Burrough, 2002; Shi, Cheung, & Zhu, 2003). Adding functional uncertainty analysis in the prototype will provide users with accuracy indexes and a degree of reliability or confidence for the geospatial data.

For the function of setting the location of the interest, the prototype allows users to choose a location from a map - which is convenient as maps are intuitive. However, a more precise way to set the location of interest is based on an address or the coordinates of a location. Therefore, it is better to provide the function of address or coordinate input to identify the location of interest. Such an extension could leverage any number of geocoding services, either via ArcMap or via the internet like Google Maps Geocoder or the like.

As mentioned before, a new GUI is needed to allow users to add more data (local or internet data). Meanwhile the new GUI will have a functionality to map the new application ontologies and schemas to the designed reference model if users provide the application ontologies.

Remotely sensed data are one of the most frequently used sources of spatial data. There are a variety of spatial and spectral data available ranging from SPOT, LandSat, QuickBird or ASTER data to RADARSAT and JERS data. Remotely sensed data also play an increasingly important role in the applications related to the Earth surface, especially in the applications involving data covering larger areas, like the long Canada-US border. As such, the integration of not only vector data but also RS data may be required. The prototype in the paper only takes into account vector data, while the methodology presented here is also capable of integrating RS data, future work will include such integration.

REFERENCES

Abel, D. J., Ooi, B. C., Tan, K. ., & Tan, S. H. (1998). Towards integrated geographical information processing. *International Journal of Geographical Information Science*, 12(4), 353-371.

Aditya, T., Kraak, M. J. (2007). A Search Interface for an SDI: Implementation and Evaluation of Metadata Visualization Strategies. *Transaction in GIS*, 11(3),413-435.

Africa SDI. (2009). Framework data background (Intro). Retrieved on Nov 15, 2009 from http://nsif.dla.gov.za/services/fdcover.html

Ahlqvist, O., Keukelaar, J., & Oukbir, K. (2003). Rough and fuzzy geographical data integration. *International Journal of Geographical Information Science*, 17(3), 223-234.

Allen, D. M., Schuurman, N., Deshpande, A., & Scibek, J. (2008). Data integration and standardization in cross-border hydrogeological studies: A novel approach to hydrostratigraphic model development. *Environmental Geology, 53*(7), 1441-1453.

ANZLIC. (2009). Australia spatial data infrastructure. Retrieved on 2009-10-20 from http://www.anzlic.org.au/infrastructure_ASDI.html

Aumueller, D., Do, H.H., Massmann, S., Rahm, E. (2005). Schema and ontology matching with COMA++. *ACM SIGMOD International Conference on Management of Data*, 2005, 906–908.

Aylesworth-Spink, S. (2009). Tradeing with Canada as partners: How any purchase American plan risks trading markets in Canada. Retrieved on 2009-11-30 from http://free-trade.suite101.com/article.cfm/buy_american_economic_stimulus_package

Balley, M., Parent,C., Spaccapietra, S. (2004). Modelling geographic data with multiple representations. *International Journal of Geographical Information Science*, 18(4), 327-352.

Bererra, R., 1993, Architecture for a federated GIS. *Research Report, Intergraph Corporation,Mapping Science Division.*

Beven, K., 1995. Linking parameters across scales: subgrid parameterizations and scale dependent hydrological models. (1995). *Hydrological Processes*, 9(5-6),507-525

Bishr, Y. (1998). Overcoming the semantic and other barriers to GIS interoperability.

International Journal of Geographical Information Science, 12(4), 299-314.

Brodeur, J., Edwards, G., Bédard, Y., & Moulin, B. (2003). Revisiting the concept of geospatial data interoperability within the scope of human communication processes. *Transactions in GIS, 7*(2), 243-265.

Brisaboa,N.R., Cotelo-Lema,J.A.,Farina,A.,Luaces,M.R.,Parama, J.R.,and Viqueira,J.R.R. (2007). Collecting and publishing large multiscale geographic datasets. *Software.practices and experience*. 37,1319-1348.

Budak, I.A., Sheth, A., Ramakrishnan, C., Lynn U.E., Usery E., Azami, M., Kwan, M.-P. (2006). Geospatial Ontology Development and Semantic Analytics. *Transactions in GIS*,10(4),551-574.

Buccella, A., Cechich,A., Gendami,D.m Lanubile,F., Semeraro, G., Colagrossi,A., (2010). GeoMergeP: Geographic information integration through enriched ontology matching. *New generation computing*. 28(1), 41-47.

Buogo, A., & Chevallier, J. -. (1995). Spatial information systems and information integration. *Computers, Environment & Urban Systems*, 19(3), 161-170.

Chen, J. (1999). Recent progress and future directions of national spatial data infrastructure in China. *Jurnal of Remote Sensing*, 3(2).

Chrisman, N. R. (1991). The error component in spatial data. *Geographical Information Systems.* Vol. 1: Principles, 165-174.

Commission of the European Communities. (2006). Communication from the commission on a European programme for critical infrastructure protection. Retrieved on 2009-11-24 from http://eur-lex.europa.eu/LexUriServ/site/en/com/2006/com2006_0786en01.pdf

Goh, C.H. (1997). Representing and reseaning about semantic conflicts in heterogeneous information sources. PhD. Thesis, MIT.

Comber, A., Fisher, P., & Wadsworth, R. (2004). Integrating land-cover data with different ontologies: Identifying change from inconsistency. *International Journal of Geographical Information Science, 18*(7), 691-708.

Crosetto, M., & Tarantola, S. (2001). Uncertainty and sensitivity analysis: Tools for GIS-based model implementation. *International Journal of Geographical Information Science*, 15(5), 415-437.

Cruz, I. F., Sunna, W., Makar, N., & Bathala, S. (2007). A visual tool for ontology alignment to enable geospatial interoperability. *Journal of Visual Languages and Computing, 18*(3), 230-254.

Deschamps, A., Greenlee, D., Pultz, T.J., and Saper, R., (2002). Geospatial data integration for application in flood prediction and management in the Red River Basin. *IEEE International Geoscience and Remote Sensing Symposium. 24th Canadian Symposium on Remote Sensing. Proceedings (Cat. No.02CH37380)*, p 3338-40 vol.6.

Devogele, T., Trevisan, J. and Raynal, L. (1996): Building a Multi-Scale Database with Scale-Transition Relationships. In: Kraak, M.J. and Molenaar, M. (eds.): *Advances in GIS research II* (Proceedings 7th International Symposium on Spatial Data Handling), London: Taylor & Francis: 6.19–6.33

Devogele, T., Parent, C., & Spaccapietra, S. (1998). On spatial database integration. *International Journal of Geographical Information Science*, 12(4), 335-352.

Donaubauer,A., Fichtinger,A., Schilcher,M., Straub,F., (2006). Model driven approach for accessing distributed spatial data using web services – Demonstrated for cross-border GIS application. Shapping the Change, XXIII FIG congress, Munich, Germany, October 8-13,2006.

Eisenberg, D.J., (2002). SVG Essentials, O'Reilly & Associates, Inc., Sebastopol, CA, 2002

ESRI, (2000). Map generalization in GIS: practical solutions with workstation ArcInfo software. Retrieved on March 10, 2010 from http://downloads2.esri.com/support/whitepapers/ao_/Map_Generalization.pdf

ESRI. (2008). ArcGIS 9.3 desktop help.

Euzenat, J., Shvaiko, P., (2007). Ontology Matching. Springer, Berlin, Heidelberg, DE, 341pp.

The FGDC. (2008a). NSDI framework. Retrieved on Nov. 15, 2009 from http://www.fgdc.gov/framework/

The FGDC. (2008b). Department of Homeland Security Geospatial Data Model. Retrieved 2008-4-12 from http://www.fgdc.gov/participation/working-groups-subcommittees/hswg/dhs-gdm/.

The FGDC. (2008c). DHS GDM v2.0- Downloadable interactive HTML data model documentation. Retrieved 2008-4-12 from http://www.fgdc.gov/participation/working-groups-subcommittees/hswg/dhs-gdm/version-2-0.

The FGDC. (2008d). Geographic information framework data standard. Retrieved on 2008-5-10 from http://www.fgdc.gov/standards/standards_publications/

The FGDC. (2009). Department of homeland security geospatial data model. Retrieved 2009-10-1 from http://www.fgdc.gov/participation/working-groups-subcommittees/hswg/dhs-gdm/.

Fileto, R. (2001). Issues on interoperability and integration of heterogeneous Geographical Data. *Proceedings of the Third Brazilian Symposium.*

Fonseca, F.T., Egenhofer, M.G., (1999). Ontology-driven geographic information systems. *Proceedings of the 7th ACM international symposium onadvances in geographic information systems, Kansas, Missouri, United States.* 1999

Fonseca, F., Davis, C., & Câmara, G. (2003). Bridging ontologies and conceptual schemas in geographic information integration. *GeoInformatica,* 7(4), 355-378.

Fonseca, F. T., Egenhofer, M. J., Agouris, P., & Cmara, G. (2002). Using ontologies for Integrated geographic information systems. *Transactions in GIS,* 6(3), 231-257.

Fonseca, F., Egenhofer,M., Davis,C., and Camara, G., (2002). Semantic granularity in ontology-driven geographic information systems. *AMAI annals of mathematics and artificial integgligence-special issue on spatial and temporal granularity* 36(1-2): 121-151.

Gao,S., Mioc,D., Anton,F., Yi,X.L.,Coleman,D.J. (2008). Online GIS services for mapping and sharing disease information. *International Journal of Health Geograhpics,* 7(8),

Gärtner, H., Bergmann, A., Schmidt, J. (2001). Object-oreinted modeling of data sources as a tool for the integration of heterogeneous geoscienctific information. *Computers and Geosciences,* 27(8),975-985.

Geobase. (2007a) National road network: conceptual data model segmented view,edition 2.0. Retrieved on 2009-08-12 from www.geobase.ca.

Geobase. (2007b) National road network: Feature catalogue segmented view, edition 2.0. Retrieved on 2009-08-12 from www.geobase.ca.

Geoconncetions, (2006). GeoConnections and U.S. FGDC streamline cross-border geospatial data sharing. Retrieved 2009-10-01, from http://www.geoconnections.org/en/aboutGeo/successStories/id=1000

Geoconnections.(2008a). *CDGI framework data definition.* Retrieved 2008-04-25, from http://www.geoconnections.org/opportunities/rfp/framework/rfp_sd_definition_E.pdf .

Geoconnections. (2008b). Cross-border content and services for critical infrastructure identification. Retrieved on 2009-11-24 from http://www.geoconnections.org/en/aboutGeo/projects/id=516

Guan, J. -., Zhou, S. -., Chen, J. -., Chen, X. -., An, Y., Yu, W., et al. (2003). Ontology-based GML schema matching for spatial information integration. *2003 International Conference on Machine Learning and Cybernetics,* Xi'an. , *4* 2240-2245.

Gugliotta, A., Davies,R., Gutierrea-billarias, L., Tanasescu,V., Domingue,J., and etc. (2006). Enhancing data and processes integration and interoperability in Emergency situations : a SWS based emergency mangement system. *Proceedings of the first international workshop on application and business aspects of the semantic web (Sebiz).* Retrieved 2008-5-8 from http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-226/sebiz2006-proceedings.pdf#page=8.

Hakimpour,F., Timpf,S. (2001). Using ontologies for resolution of semantic heterogeneity in GIS. *The 4th AGILE Conference on Geographic Information Science* in Brno.

Hakimpour, F. (2003). Using ontologies to resolve semantic heterogeneity for integrating spatial database schemata. Ph.D. Dissertation, Zurich University, Switzerland.

Hakimpour, F., Geppert, A. (2002). Global schema generation using formal ontologies. *Proceedings of the ER'02: 21st International Conference on Conceptual Modeling. Lecture Notes in Computer Science*, vol. 2503. Springer, Berlin, Heidelberg, pp. 307–321.

Hart, G., Dolbear, C., and Goodwin, J., (2007). Lege Feliciter: Using Structured English to represent a Topographic Hydrology Ontology. *Proceedings of the OWL Experiences and Directions Workshop*, 2007, Retrieved on August 22, 2008 at http://ncg.nuim.ie/gisruk/materials/proceedings/PDF/2A5.pdf.

Heuvelink, G. B. M. (1998). Uncertainty analysis in environmental modelling under a change of spatial scale. *Nutrient Cycling in Agroecosystems, 50*(1-3), 255-264.

Heuvelink, G. B. M. (1999). Propagation of error in spatial modeling with GIS. *Geographical Information Systems: Principles, Techniques, Applications and Management*, 188-192.

Heuvelink, G. B. M., & Burrough, P. A. (2002). Developments in statistical approaches to spatial uncertainty and its propagation. *International Journal of Geographical Information Science*, 16(2), 111-113.

Homeland Security. (2008). Homeland Security Presidential Directive 7: Critical infrastructure identification, prioritization, and protection. Retrieved on 200-11-24 from http://www.dhs.gov/xabout/laws/gc_1214597989952.shtm

International Boundary Commissioin. The international boundary. Retrieved 2009-10-1, from http://www.internationalboundarycommission.org/boundary.html

Izza,S. (2009). Integration of industrial information systems: from syntactic to semantic integration approaches. *Enterprise Information Systems*, 3(1), 1-57.

Jaenicke,K., (2003). Set-up of a cross-border spatial data infrastructure for an interoperable use of data through the internet. *ESRI User Conference 2003*,07-11 July 2003,San Diego.

Jos de Bruijn. (2003). Using ontologies - enabling knowledge sharing and reuse on the semantic web. *Technical Report DERI-2003-10-29, DERI.*

Jones,C.B., Kidner, D.B., Luo,L.A., Bundy,G.L., (1996). Database design for a mutli-spatial information system. *International journal of geographical information sstems.* 10(8),901-920.

Kalfoglou, Y. and Schorlemmer M. (2003). Ontology mapping: the state of the art, *The Knowledge Engineering Review* 18 (1) (2003), pp. 1–31.

Karla A.C. Borges, Clodoveu A. Davis and Alberto H.F. Laender (2001), Omt-g: an object-oriented data model for geographic applications, *Geoinformatica* , 5 (3), 221–260.

Kavouras, M., Kokla, M., Tomai, E., (2005). Comparing categories among geographic ontologies. *Computer & Geosciences.* 31(2005), 145-154.

Klopfer, M. (2005). Interoperability & Open architectures: an analysis of existing standardisation processes & procedures. Retrieved 2008-04-18, from http://www.opengeospatial.org/pressroom/papers.

Klein, M., (2001). Combining and relating ontologies: an analysis of problems and solutions. *In: Proceedings of the IJCAI'01: 17th International Joint Conferences on Artificial Intelligence, Seattle, WA, USA*, pp. 53–62.

Kröger, W. (2008). Critical infrastructures at risk: A need for a new conceptual pproach and extended analytical tools. *Reliability Engineering and System Safety* 93 (12), pp. 1781-1787

Lehto L., Sarjakoski, T., Hvas,A., Hoolander, P., Ruotsalainen, R., Illert A. (2004). A prototype cross-border GML data service. *7th AGILE Conference on Geographic Information Science, 29 April-1May 2004, Heraklion.*

Lilburne, L.R., Webb, T.H., Benwell, G.L., 2004. The scale matcher:a procedure for assessing scale compatibility of spatial data and models. *Int. J. Geogr. Inform. Sci.* 18, 257–279.

Lin,K., Ludäscher, B. (2006). A system for semantic integration of geologic maps via ontologies. *Semantic web technologies for searching and retrieving scientific data (SCISW), 2006.*

Lutz, M., & Klien, E. (2006). Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science*, 20(3), 233-260.

Madhavan, J., Bernstein, P., Rahm, E. (2001). Generic Schema Matching with Cupid. Retrieved 2008-4-30 from http://research.microsoft.com/~philbe/CupidVLDB01.pdf

Maedche A., and Staab S., (2002) . Measuring similarity between ontologies. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 251–263, 2002.

Masser, I. (2000). Spatial data infrastructures in Europe. *Proceedings of the International Conference—Quo Vadis Surveying of the 21st Century, FIG Working Week 2000 in Prague, Czech Republic*, 22-26 May 2000.

Masser,I. (2003). An overview of SDI activities. Retrieved on 2009-10-30 from http://www.gsdidocs.org/gsdiconf/GSDI-7/papers/TU-1A.pdf

McBride, B., (2002). Jena: A semantic web toolkit. *IEEE Internet Computer.* 6(6), 55-58.

Microsoft. (2009). Access 2003 help and how-to. Retrieved on 2009-11-10 from http://office.microsoft.com/en-us/access/fx100646921033.aspx

Moreno-Sanchez, R., Anderson, G., Cruz, J., & Hayden, M. (2007). The potential for the use of open source software and open specifications in creating web-based cross-border health spatial information systems. *International Journal of Geographical Information Science, 21*(10), 1135-1163.

Mustiere,S., (2005) Cartographic generalization of roads in a local and adaptive approach: A knowledge acquisition problem. *International Journal of Geographical information systems.* 19(8-9),937-955.

The OGC. (2006). *Geospatial Semantic Web* Interoperability Experiment Report. Retrieved 2009-10-1 from http://www.opengeospatial.org/projects/initiatives/gswie

The OGC. (2008). http://www.opengeospatial.org/standards. Retrieved 2008-4-10 from http://www.opengeospatial.org/standards.

The OGC. (2009). The OpenGIS abstraction specification Topic 5: Features. Retrieved 2009-10-1, from http://www.opengeospatial.org/standards/as

Parent, C., Spaccapietra, S., Zimányi, E., (1999). Spatio-temporal conceptual models: data structures+space+time. *In: Proceedings of the GIS '99: 7th ACM International Symposium on Advances in Geographic Information Systems. ACM Press, New York, NY, USA*, pp. 26–33.

Parent,C. and Spaccapietra,S. (2000). Database integration: the key to data interoperability. In Papazoglou,M., Spaccapietra,S. Tari,Z. (Ed.), *Advances in object-oriented data modeling* (pp. 221-254). Cambridge, Mass. [u.a.] : MIT Press.

Parent C., Spaccapietra S., and Zimányi E. (2005). The murmur project: modeling and querying multi-representation spatio-temporal databases, *Information Systems* **31** (8), 733–769

Peachavanish, R., Karimi, H.A. (2007). Ontological Engineering for Interpreting Geospatial Queries. *Transactions in GIS*, 11(1), 115-130.

Peng,Zh. R., and Zhang, Ch,R., (2004). The role of geography markup language (GML), scalable vector graphics (SVG), and Web feature service (WFS) specifications in the development of Internet geographic information systems (GIS*). Journal of Geographic Systems*. 6(2). 95-116.

Petersen, K.L. (2008). Terrorism: when risk meets security. *Alternatives*, 33 (2008), 173-190.

Public Safety Canada. (2008). Working towards a national strategy and action plan for critical infrastructure. Retrieved on 2009-11-24 from http://www.publicsafety.gc.ca/prg/em/cip/_fl/nat-strat-critical-infrastructure-eng.pdf

Public Safety Canada. (2009). Critical infrastructure protection. Retrieved on 2009-11-24 from http://www.publicsafety.gc.ca/prg/em/cip-eng.aspx.

Pundt, H., & Bishr, Y. (2002). Domain ontologies for data sharing - an example from environmental monitoring using field GIS. *Computers and Geosciences, 28*(1), 95-102.

Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal, 10*(4), 334-350.

Riecken, J., Bernard,L., Portele,C., Remke,A. (2003). North-Rhine Westphalia: Building a regional SDI in a cross-border environment ad-hoc integration of SDIs: lessons learnt. *9th EC-GI & GIS Workshop.* 25-27 June 2003, A Coruña

Rodriguez,A., Egenhofer,M., Rugg,R., (1999). Assessing semantic similarities among geospatial feature class definitions. *In: Interoperating geographic information systems, second international conference, INTEROP'99, Zurich, Switzerland.*

Rodriguez,M.,Egenhofer,M.,(2004).Comparinggeospatialentityclasses:an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science* 18(3), 229–256.

Safra, E., kanza Y., Sagiv, Y., Beeri, C., Doytsher, Y., (2010) Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets. *International journal of geographical information science.* 24(1), 69-106.

Samal, A., Seth, S., and Cueto, K., (2004). A feature-based approach to conflation of geospatial sources. *International journal of geographical information science,* 18(5), 459-489.

Schwering, A., Raubal, M., (2005). Spatial relations for semantic similarity measurement. *In: Proceedings of the ER'05: 24th International Conference on Conceptual Modeling. Lecture Notes in Computer Science*, vol. 3770. Springer, Berlin, Heidelberg, pp. 259–269.

Sheth, A. P., & Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *Computing Surveys, 22*(3), 183-236.

Shi, W., Cheung, C. K., & Zhu, C. (2003). Modelling error propagation in vector-based buffer analysis. *International Journal of Geographical Information Science*, 17(3), 251-271.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., Katz, Y., (2007). Pellet: A practical OWL-DL reasoned. *Web Semantics* 5(2), 51-53.

Sotnykova, A., Vangenot, C., Cullot, N., Bennacer, N., Aufaure, M., (2005). Semantic mappings in description logics for spatio-temporal data- base schema integration. *Journal on Data Semantics III*, 143–167.

Souza, D., Salgado, A. C., & Tedesco, P. (2006). Towards a context ontology for geospatial data integration. *OTM 2006 Workshops - OTM Confederated International Workshops, Montpellier.* , 4278 LNCS - II 1576-1585.

Stoimenov,L. , Djordjevic-Kajan,S. (2005). An architecture for interoperable GIS use in a local community environment. *Computers and Geosciences*, 31(2005), 211-220.

Stoimenov, L., Stanimirovic, A., Djordjevic-Kajan, S. (2006). Discovering mappings between ontologies in semantic integration process. *Proceedings of the AGILE'06: 9th Conference on Geographic Information Science*, Visegrád, Hungary, pp. 213–219.

The Carbon Poject. (2009). Cross-border content and services for critical infrastructure identification. Retrieved 2009-10-1 from http://www.fgdc.gov/grants/2008CAP/Reports/134-08-4-MA-InterimReport.pdf

The Senate of the United States. (2001). H.R.3162, the USA PATRIOT Act of 2001, An Act to deter and punish terrorist acts in the United States and around the world, to enhance law enforcement investigatory tools, and for other purpose. Retrieved on 200-11-24 from http://epic.org/privacy/terrorism/hr3162.pdf.

The White House Washington. (1998). Presidential directive PDD-63. Retrieved on 2009-11-24 from http://www.fas.org/irp/offdocs/pdd/pdd-63.htm.

Tolone,W.J., Wilson,D., Raja,A., Xiao,W.N. and *et al.* (2004). Critical infrastructure integration modeling and simulation. *Lecture Notes in Computer Science* 3073, pp. 214-225

Tomai, E., Prastacos, P., Kavouras, M. (2007). A Framework for Intensional and Extensional Integration of Geographic Ontologies. *Transactions in GIS*, 11(6), 873-887.

Uitermark, H., van Oosterom, P., Mars, N., & Molenaar, M. (1999). Ontology-based geographic data set integration. *Spatio-Temporal Database Management,*, 60-78.

Uitermark, H.T. (2001). Ontology-based geographic data set integration. PhD Thesis. University of Twente, Enschede, The Netherlands (http:// www.ub.utwente.nl/ webdocs/ inf/ 1/ t000001e.pdf).

Uitermark, H. T., van Oosterom, P. J. M., Mars, N. J. I., & Molenaar, M. (2005). Ontology-based integration of topographic data sets. *International Journal of Applied Earth Observation and Geoinformation*, 7(2), 97-106.

U.S. Census Bureau. (2008). TIGER/Line Shapefiles technical documentation. Retrieved on 2008-08-20 from http://www.census.gov/geo/www/tiger/tgrshp2008/tgrshp2008.html

Varanka, D., (2008). National Topographic Modeling, Ontology-Driven Geographic Information in the Context of *The National Map: First International Workshop on Information Semantics and its Implications for Geographical Analysis (ISGA '08) at GIScience 2008, the 5th International Conference on Geographic Information Science, Park City, Utah,* September 23, 2008.

Visser, U., Stuckenschmidt, H., Schuster, G., & Vögele, T. (2002). Ontologies for geographic information processing. *Computers and Geosciences, 28*(1), 103-117.

Visser, U. (2004). Intelligent information integration for the semantic web. *Lecture Notes* *in Computer Science*, vol. 3159. Springer, Berlin, Heidelberg, 150pp.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S., (2001). Ontology-based integration of information—a survey of existing approaches. *Proceedings of the IJCAI'01: 17th International Joint Conferences on Artificial Intelligence, Seattle, WA*, pp. 108–117.

Walter, V., & Fritsch, D. (1999). Matching spatial data sets: A statistical approach. *International Journal of Geographical Information Science*, 13(5), 445-473.

Wang, Y.D., Gong, J.Y., Wu,X.H. (2007). *Geo-spatial Information Science*. 10(3). 204-207

Wang, Y.H., Wei,F.Y.(2008). A schema-matching-based approach to propagating Updates between heterogeneous spatial databases. *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Advanced Spatial Data Models and Analyses.*

Wikipedia. (2006). List of Canada-United States border crossing. Retrieved on 2009-10-20 from http://en.wikipedia.org/wiki/List_of_Canada_–_United_States_border_crossings

Wikipedia. (2007). Critical infrastructure. Retrieved on 2009-11-24 from http://en.wikipedia.org/wiki/Critical_infrastructure.

Wikipedia. (2008). Northeast blackbout of 2003. Retrieved on 2009-11-29 from http://en.wikipedia.org/wiki/Northeast_Blackout_of_2003#Sequence_of_events.

www.nsif.dla.gov.za. (2009). Welcome to national spatial information framework. Retrieved on Nov. 15, 2009, from www.nsif.dla.gov.za.

Wolthusen, S. D. (2005). GIS-based command and control infrastructure for critical infrastructure protection. Retrieved on 2009-11-24 from http://portal.acm.org/citation.cfm?id=1108122

Yunpeng, Y., Jiancun, L., & Zhengmin, H. (2006). Research on ontology based semantic integration model in spatial data sharing. *2006 IEEE International Geoscience and Remote Sensing Symposium, IGARSS, Denver, CO*. 2872-2875.

Zhang,D., Peng,Z., Li, W., Day,M.J. (2003). GML-Based Interoperable Geographical Databases. *Cartography*, 32 (2), 1-16.

Zhang, Z., (2005). Ontology query languages for the semantic web: a performance evaluation. M.Sc. Thesis, University of Georgia, Athens, 158pp.

Zhou, J., Guan, J., Zhu, F., Zhou, S., & Li, P. (2006). Ontology-based semantic integration of GML spatial information. *Geoinformatics 2006: Geospatial* Information Science, Wuhan. , 6420.

APPENDIX

CODES

The following codes are the primary codes in the prototype. The function of **Intersect** is to clip the specified DOI data based on a buffer area defined by the locations of WOI and the inputted buffer distance. The function of **Subframeworkdata_interset(subframeworkdataname As String, buff_feat As String)** is to revise attribute data based on the developed ontology database and merger all the retrieved data based on data geometry types.

```
Function intersect()
   Dim GP As IGeoProcessor
   Dim pGpEnumList As IGpEnumList
   Dim pGPextend As IGPExtent
   Dim buff_feat As String
   Dim out_feat As String
   Dim input_feat As String
   Dim parameters As IVariantArray
   Dim i As Integer
   Dim j As Integer
   Dim k As Integer
   Dim count0 As Integer
   Dim count1 As Integer
   Dim sFCname As String
   Dim shapefilenamelist As String

   Set GP = New GeoProcessor
   GP.OverwriteOutput = True
   'GP.AddOutputsToMap = True
```

```
 GP.SetEnvironmentValue "workspace", filepath & "\inputdata"

 buff_feat = filepath & "\outputdata\output.shp"

Dim node As node
Dim nodex As node
Dim t As New subFrameworkdata
Dim t1 As New myframeworkdata
Dim flag As Boolean

For i = 1 To TreeView1.Nodes.Count 'for each data topic and sub data topic
    Set node = TreeView1.Nodes(i)
   If node.Children = 0 And node.Checked = True Then 'for no child node
     subframeworkdata_interset node.Text, buff_feat
    End If
Next

End Function

Sub Subframeworkdata_interset(subframeworkdataname As String, buff_feat As String)
  Dim GP As IGeoProcessor
 Dim out_feat As String
 Dim input_feat As String
 Dim parameters As IVariantArray
 Dim i As Integer
 Dim j As Integer
 Dim k As Integer
 Dim count0 As Integer
 Dim count1 As Integer ' for the result shapefile
 Dim shapefilenamelist(2) As String
 Dim pGpUtils As IGPUtilities
```

```
Dim ttt As String
Dim pTargetFeatureLayer As IFeatureLayer
 Dim cCurrentRefField() As New refField
Dim cc9 As Integer
'step 1: to get the shapefile for each sub frameworkdata
 Set rs = New ADODB.RecordSet
rs.Open "select * from " & subframeworkdataname & "_shapefile", db, adOpenStatic,
adLockOptimistic
count0 = 0
'if there is no record in ***_shapefile table which means no data collected for it, so no
need to do the further process.
If rs Is Nothing Then
  MsgBox "no data, occured in subframeworkdata_interset()"
End If
 'get the shapefile
Do
   ReDim Preserve cShapefile0(count0)
   Set cShapefile0(count0) = New shapefile
   cShapefile0(count0).shapefileName = rs.fields("shapefilename").value
   cShapefile0(count0).country = rs.fields("country").value
   rs.MoveNext
   count0 = count0 + 1

Loop Until rs.EOF
rs.Close
Set rs = Nothing
count0 = 0
'step2: clip for canada & usa
For j = LBound(cShapefile0) To UBound(cShapefile0)
     Set parameters = New VarArray
     Set GP = New GeoProcessor
```

```
        out_feat = filepath & "\outputdata\" & cShapefile0(j).shapefileName & "c.shp"
        input_feat = filepath & "\inputdata\" & cShapefile0(j).shapefileName & ".shp"

        parameters.Add (input_feat)
        parameters.Add (buff_feat)
        parameters.Add (out_feat)
        GP.Execute "Clip_analysis", parameters, Nothing

        ReDim Preserve resultShapefileCAN(count0)
        Set resultShapefileCAN(count0) = New shapefile
        resultShapefileCAN(count0).shapefileName = cShapefile0(j).shapefileName &
"c.shp"
       ' resultShapefileCAN(count0).country = "Canada"
        count0 = count0 + 1

        Set parameters = Nothing
        Set GP = Nothing
        Set pGpUtils = New GPUtilities
        pGpUtils.ReleaseInternals
  Next

  'step3: the next step after clip should be to revise the attribute to the standard one then
to merge it,
  'otherwise the shapefile cannot be merged.
 'first create a template for the data topic, need to determine if the shapefile is already
exist.
  'to get the standard shapefile field name and datatype which will be used to create the
template.
   count0 = 0
   Set rs = New ADODB.RecordSet
```

```
  rs.Open "select * from " & subframeworkdataname, db, adOpenStatic,
adLockOptimistic
  Do
    ReDim Preserve shapefileFields(count0)
    Set shapefileFields(count0) = New myfield
    shapefileFields(count0).name = rs.fields("field").value
    shapefileFields(count0).ontology = rs.fields("ontology").value
    shapefileFields(count0).fieldtype = rs.fields("fieldtype").value
    count0 = count0 + 1
    rs.MoveNext
  Loop Until rs.EOF
  rs.Close
  Set rs = Nothing




 'add the canada data into canada template.
  'open the shapefile from file name, select all the records and then get the attribute value
 Dim pWorkspaceFactory As IWorkspaceFactory
 Dim pFeatureWorkspace As IFeatureWorkspace
 Dim pFeatureLayer As IFeatureLayer
 Dim psourceFeatClass As IFeatureClass
 'Create a new ShapefileWorkspaceFactory object and open a shapefile folder
 Set pWorkspaceFactory = New ShapefileWorkspaceFactory
 Set pFeatureWorkspace = pWorkspaceFactory.OpenFromFile(filepath & "\outputdata",
0)
 Dim pfields As IFields
 Dim pcursor As IFeatureCursor
 Dim tt As String
 Dim t As String
```

```
Dim flag11 As Boolean
Dim flag22 As Boolean
Dim pFieldEdit As IFieldEdit
Dim pFeat As IFeature
Dim pnewFeat As IFeature
Dim pfield As IField
Dim ptemField As IField
Dim pTargetField As IField
'Dim pRow As IRow
Dim fieldIndex As Long
Dim targetFieldIndex As Long
Dim rs1 As ADODB.RecordSet
Dim rs2 As ADODB.RecordSet
Dim m As Integer
Dim tempvalue As String
Dim clipReviseresult() As String
Dim cc0 As Integer

Dim pGP As Object
Dim ptable As ITable
Dim pnewfieldEdit As IFieldEdit
Dim pFL As IFeatureLayer

'for each data source get the semantic relation between the application concepts and standard concept.
For i = LBound(resultShapefileCAN) To UBound(resultShapefileCAN) 'resultshapefilecan for the clipped shapefile of canada like canada_provincec.shp,canada_province2c.shp
  t = Left(resultShapefileCAN(i).shapefileName, Len(resultShapefileCAN(i).shapefileName) - 4) 't is the clipped shapefile
  Set psourceFeatClass = Nothing
```

```
Set psourceFeatClass = pFeatureWorkspace.OpenFeatureClass(t)
Set pFL = New FeatureLayer
Set pFL.FeatureClass = psourceFeatClass
pFL.name = psourceFeatClass.AliasName
 'if there is no features in the shapefile then goto the next shapefiel
 If psourceFeatClass.FeatureCount(Nothing) = 0 Then
 GoTo hh4
 End If
 flag11 = False

 tt = Left(t, Len(t) - 1) ' the orginal shapefile name before clip 'tt is the original
shapefile
 count0 = 0
 ReDim cCurrentRefField(0)
 For j = LBound(shapefileFields) To UBound(shapefileFields)
    Set rs = New ADODB.RecordSet
    ttt = "select * from " & subframeworkdataname & "_" & shapefileFields(j).name &
" where tablename='" & tt & "'"
    rs.Open ttt, db, adOpenStatic, adLockOptimistic
    If rs.EOF And rs.BOF Then
      ReDim Preserve cCurrentRefField(count0) 'even thought there is no table for an
integrated field, still need to create it.
      With cCurrentRefField(count0)
              .fieldname = ""
              .ontology = ""
              .sRelation = ""
              .value = ""
              .StandardField = shapefileFields(j).name
              .fieldtype = shapefileFields(j).fieldtype
           End With
           count0 = count0 + 1
```

```
  GoTo hh0
  Else
  rs.MoveFirst
 End If
'if the shapefile has the field,record it
Do

    ReDim Preserve cCurrentRefField(count0)
      If rs.fields("value").value = "has" Then
        Set rs1 = New ADODB.RecordSet
        'rs1 is the standard road classes
        rs1.Open "select * from " & subframeworkdataname & "_" & shapefileFields(j).name & "_value", db, adOpenStatic, adLockOptimistic
        If rs1.EOF And rs1.BOF Then
          GoTo hh2
          Else
          rs1.MoveFirst
        End If
        Do
          Set rs2 = New ADODB.RecordSet
          rs2.Open "select * from " & subframeworkdataname & "_" & shapefileFields(j).name & "_value_" & rs1.fields("field").value & " where tablename='" & tt & "'", db, adOpenStatic, adLockOptimistic
          If rs2.EOF And rs2.BOF Then
          GoTo hh3
          Else
          rs2.MoveFirst
          End If
           Do
              ReDim Preserve cCurrentRefField(count0)
              With cCurrentRefField(count0)
```

```
                .fieldname = rs2.fields("field").value
                .ontology = IsNull(rs2.fields("ontology").value)
                .sRelation = IsNull(rs2.fields("relation").value)
                .value = rs2.fields("value").value
                .StandardField = shapefileFields(j).name
                .StandardField2 = rs1.fields("field").value
                .fieldtype = shapefileFields(j).fieldtype
            End With
            count0 = count0 + 1
          rs2.MoveNext
        Loop Until rs2.EOF
        rs2.Close
        Set rs2 = Nothing
hh3:
      rs1.MoveNext
      Loop Until rs1.EOF
      rs1.Close
      Set rs1 = Nothing
    Else
      With cCurrentRefField(count0)
        .fieldname = rs.fields("fieldname").value
        .ontology = IsNull(rs.fields("ontology").value)
        .sRelation = rs.fields("relation").value
        .value = IsNull(rs.fields("value").value)
        .StandardField = shapefileFields(j).name
        .fieldtype = shapefileFields(j).fieldtype
      End With
      count0 = count0 + 1
      GoTo hh0
    End If
hh2:
```

```
            rs.MoveNext
        Loop Until rs.EOF
        rs.Close
        Set rs = Nothing

hh0:
    Next j

        'directed revise the field value and field name
        ' open the database
        'get the field, just create a new field with the integrated name
        'and then if field=field , copy the value row by row
        ' or if field=field value, get the value of each row,
        ' at last delete the old field
       Set ptable = psourceFeatClass
     Set pGP = CreateObject("esriGeoprocessing.GPDispatch.1")
     pGP.AddToolbox "C:/Program Files/ArcGIS/ArcToolbox/Toolboxes/Data Management Tools.tbx" 'is this useful or not?
    For m = LBound(cCurrentRefField) To UBound(cCurrentRefField)
        If cCurrentRefField(m).StandardField <> "" Then   'means the table is not created,but still need to create the field
            If cCurrentRefField(m).StandardField2 = "" Then
              If cCurrentRefField(m).fieldname <> cCurrentRefField(m).StandardField Then  'if the names are different
                pGP.addfield_management ptable, UCase(cCurrentRefField(m).StandardField), UCase(cCurrentRefField(m).fieldtype)
              End If
              If cCurrentRefField(m).fieldname <> "" And UCase(cCurrentRefField(m).fieldname) <> UCase(cCurrentRefField(m).StandardField) Then
```

```
            pGP.CalculateField_management filepath & "\outputdata\" & t & ".shp", UCase(cCurrentRefField(m).StandardField), "[" & UCase(cCurrentRefField(m).fieldname) & "]", "VB", ""
          End If
        Else
          If flag11 = False Then
            pGP.addfield_management ptable, UCase(cCurrentRefField(m).StandardField), cCurrentRefField(m).fieldtype
            flag11 = True
          End If
'
          pGP.Selectlayerbyattribute_management pFL, "NEW_SELECTION", cCurrentRefField(m).fieldname & " = " & Chr(39) & cCurrentRefField(m).value & Chr(39)
          pGP.CalculateField_management pFL, UCase(cCurrentRefField(m).StandardField), Chr(34) & cCurrentRefField(m).StandardField2 & Chr(34), "VB", ""
        End If
      End If
  Next

'delete non-standard fields
 m = 0
 Set pfields = psourceFeatClass.fields
 Do While m < pfields.FieldCount
   Set pfield = pfields.field(m)
   flag11 = False
   For k = LBound(cCurrentRefField) To UBound(cCurrentRefField)
      If pfield.name = UCase(cCurrentRefField(k).StandardField) Then
       flag11 = True
       Exit For
```

```
      End If
    Next

    If pfield.name = "FID" Or pfield.name = "ObjectID" Or pfield.name = "Shape" Then
    m = m + 1
    Else
      If flag11 = False Then
       psourceFeatClass.DeleteField pfield
       Else
       m = m + 1
      End If
    End If

 Loop

 ReDim Preserve clipReviseresult(cc9)
 clipReviseresult(cc9) = resultShapefileCAN(i).shapefileName
 cc9 = cc9 + 1
hh4:
 Next i

    'merge the clipped and revised datasource named tt & "_clip" where tt = Left(resultShapefileCAN(i).shapefileName, Len(resultShapefileCAN(i).shapefileName) - 5)
  For i = LBound(clipReviseresult) To UBound(clipReviseresult)
   Set psourceFeatClass = Nothing
   Set psourceFeatClass = pFeatureWorkspace.OpenFeatureClass(clipReviseresult(i))
    pShapeGeometryType = psourceFeatClass.ShapeType

  '  for different geometry types
   If pShapeGeometryType = 3 Then 'polyline
```

```
      If shapefilenamelist(0) = "" Then
          shapefilenamelist(0) = filepath & "\outputdata\" & clipReviseresult(i) & "; "
      Else
          shapefilenamelist(0) = shapefilenamelist(0) & filepath & "\outputdata\" &
clipReviseresult(i) & ";
      End If
   End If

   If pShapeGeometryType = 4 Then 'polygon
      If shapefilenamelist(1) = "" Then
          shapefilenamelist(1) = filepath & "\outputdata\" & clipReviseresult(i) & "; "
      Else
          shapefilenamelist(1) = shapefilenamelist(1) & filepath & "\outputdata\" &
clipReviseresult(i) & "; "
      End If
  End If
   If pShapeGeometryType = 1 Then 'point
      If shapefilenamelist(2) = "" Then
             shapefilenamelist(2) = filepath & "\outputdata\" & clipReviseresult(i) & "; "
         Else
             shapefilenamelist(2) = shapefilenamelist(2) & filepath & "\outputdata\" &
clipReviseresult(i) & "; "
         End If
   End If
Next i

  Dim tempname As String

 For i = LBound(shapefilenamelist) To UBound(shapefilenamelist)
   If shapefilenamelist(i) <> "" Then
     shapefilenamelist(i) = Left(shapefilenamelist(i), Len(shapefilenamelist(i)) - 1)
```

```
Set parameters = Nothing
Set GP = Nothing
Set parameters = New VarArray
Set GP = New GeoProcessor
parameters.Add (shapefilenamelist(i))
If i = 0 Then
 parameters.Add (filepath & "\outputdata\" & subframeworkdataname & "_polyline"
& ".shp")
 tempname = subframeworkdataname & "_polyline"
End If
If i = 1 Then
parameters.Add (filepath & "\outputdata\" & subframeworkdataname & "_polygon"
& ".shp")
tempname = subframeworkdataname & "_polygon"
End If
If i = 2 Then
parameters.Add (filepath & "\outputdata\" & subframeworkdataname & "_point" &
".shp")
tempname = subframeworkdataname & "_point"
End If
GP.Execute "Merge", parameters, Nothing
Set parameters = Nothing
Set GP = Nothing
Set pGpUtils = New GPUtilities
pGpUtils.ReleaseInternals
  Set psourceFeatClass = Nothing
  Set psourceFeatClass = pFeatureWorkspace.OpenFeatureClass(tempname)
  Set pTargetFeatureLayer = New FeatureLayer
  Set pTargetFeatureLayer.FeatureClass = psourceFeatClass
  pTargetFeatureLayer.name = subframeworkdataname & i
  pmap.AddLayer pTargetFeatureLayer
```

```
    pmxdoc.activeview.PartialRefresh 32, Nothing, Nothing
  End If
 Next
End Sub
```